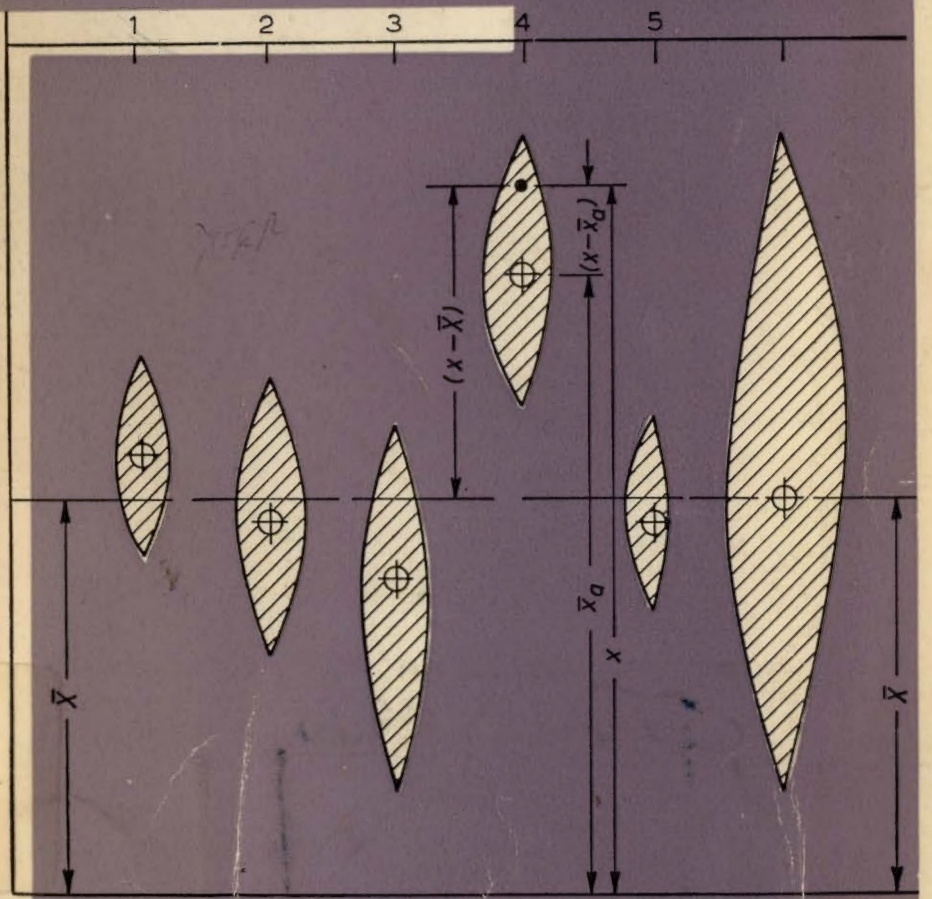
STATISTICAL METHODS FOR ENGINEERS

J. J. LEEMING



BLACKIE

LEEMING STATISTICAL METHODS FOR ENGINEERS

519.
024
62
LEE

THE LIBRARY

THE HARRIS COLLEGE

CORPORATION STREET, PRESTON.

All Books must be Returned to the College Library or Renewed not later than the last date shown below.

13 MAR 1970
17 NOV. 1970
26 MAR 1971
23 JUN 1972
13 DEC 1974
2 JUL 1976
1 JUL 197
2 APR 198
5 FEB 198

519.02462 LEE
A/C 024990
30107 000 546 314

STATISTICAL METHODS FOR ENGINEERS

STATISTICAL METHODS FOR ENGINEERS

J. J. LEEMING

B.Sc. (Oxon.), A.C.G.I., F.I.C.E., M.I.Struct.E., M.I.Mun.E., F.Inst.H.E.

BLACKIE

London and Glasgow

BLACKIE & SON LIMITED
5 FITZHARDINGE STREET
PORTMAN SQUARE
LONDON · W.1
BISHOPBRIGGS, GLASGOW

BLACKIE & SON
(INDIA) LIMITED
103-5 FORT STREET
BOMBAY

J. J. LEEMING © 1963
FIRST PUBLISHED 1963
REPRINTED 1966
REPRINTED 1969

b63136888
HARRIS COLLEGE
PRESTON
519.02762
ST912 LEE
24990
SUBE 1.70
c 20

PRINTED IN GREAT BRITAIN BY NEILL & CO. LTD., EDINBURGH

Preface

There are many books on statistical methods, but most of them seem to be written for economists and biologists, and an engineer who wishes to study the subject is to some extent handicapped by having to cope with unfamiliar terms, particularly in the examples. I have written this book with his needs, and especially those of the traffic engineer, very much in mind. Many of the examples are taken from studies on which I have been engaged during the course of my work on roads. For the traffic engineer, in particular, statistical methods are an essential tool, and he can hardly carry out his work without them.

The book is not a complete manual of the subject, and the underlying theory is outside its scope. A small bibliography is included, and some of the works in that include fuller bibliographies for further reference. The superior numbers after names in the text are references to the bibliography.

As the book is also intended for students, I have included some elementary examples, and described them in some detail. For this I can only ask the indulgence of those more advanced in the subject.

My thanks are due to Professor A. N. Black, to Professor H. E. Daniels, and especially to Dr. Alan Miller, for much helpful criticism and advice, to my wife and daughter for help with the typing of the manuscript, and to the Institutions of Civil Engineers, of Structural Engineers, of Municipal Engineers and of Highway Engineers for the use of material taken from papers presented to them.

I am specially indebted to the late Professor Sir Ronald Fisher, F.R.S., Cambridge; Dr. Frank Yates, F.R.S., Rothamstead; and Oliver & Boyd Ltd., Edinburgh, for permission to reprint Tables E to H from their book *Statistical Tables for Biological, Agricultural and Medical Research*. For more complete tables than those given here, reference should be made to that work.

J. J. LEEMING
January 1963

Contents

Tables

I TABLES IN THE MAIN TEXT

TABLES

II TABLES AT THE END OF THE BOOK

CHAPTER 1

Introductory: Terminology and Symbols

'A statistical analysis, properly conducted, is a delicate dissection of uncertainties, a surgery of suppositions.' MORONEY.[1]

'To know that no quality is sharply and uniquely defined, that every quality is of a blurred statistical nature, is of first importance to the man who represents it by a single number.'
PROFESSOR LEVY.*

The two quotations above aptly describe the fundamental principles underlying statistical methods. These are used to study variations and uncertainties. The variations and uncertainties may arise from ordinary experimental error, or, more probably—and especially more probably in traffic studies, where our data are of a particularly blurred statistical nature—they will arise from the essential complexity of the study itself. This complexity may be due to the interaction of many factors. Some of these may be unknown to us—indeed they may not even occur to us—while we do know others, and may include them in our study. In traffic studies, of course, we are dealing especially with that most unpredictable of all factors, human nature.

The use of statistical methods in our studies will enable us to achieve a number of useful ends. First of all, we can present a mass of apparently indigestible data in the form of more easily understood 'parameters', or representative figures, or by means of tables or diagrams, and so present them to our employers, or the public, or their representatives such as councillors. It is, however, only fair to add that they can be presented in misleading form, and so we must be able to perceive this, and avoid it.

A second end, of particular importance in our types of study, is that we can study the interaction of the various effects. Assume that we have

* James Forrest Lecture. Institution of Civil Engineers, 1953.

three, denoted by a, b and c. It is a not uncommon practice to fix one, say c, and then study the variation of a and b. But to do this may have the effect of falsifying the result, because if c is important the variation of a and b may not be quite the same with a fixed c as it would be if c were allowed to vary freely. By the use of statistical methods we can allow all factors to vary, and then analyse the resulting data to find those which are important, and those which are not important. For example, in a study of road curvature carried out by Professor A. N. Black and myself, the behaviour of drivers on curves was a fundamental part of the study. It was obviously possible that this might vary according to the hand of the curve. If this was so, it would then clearly be necessary to analyse each hand of curve separately, and that would have doubled the work involved. In an early stage of the analysis, however, it was found that there was no reason to think that the drivers' behaviour on curves of different hand varied materially; so all curves could be included in the same analysis, without increase of work.

This leads to a third end in the use of statistical methods, the cheapening of work by the reduction of time and labour. We are faced with a double difficulty. On the one hand, the variations necessarily mean that the volume of data collected must be large, so that any extremes may cancel each other's effects. On the other hand, the large volume of data involves time and labour, and therefore cost. The judicious use of statistical methods will enable us to arrive at the best compromise between excessive cost on the one hand and inaccuracy due to insufficient data on the other.

Terminology and Symbols

There is some difference in practice in statistical terminology and symbols, especially in the latter. Some of these are of considerable, and possibly unnecessary, complexity, and some standardization is very desirable. In this book, the practice of R. A. Fisher has been followed, with one or two very minor modifications. Not only is Fisher one of the founding fathers of the subject, but his symbols are probably the most practical and simple.

In the rest of the book, when a term is defined it is shown in italics. The number of the page on which it is defined is also shown in italics in the index.

Most of the actual symbols are described below when they first occur, but some of the more general are described here. The sum of a number of quantities of a function denoted by some such symbol as x is denoted by $S(x)$. So this symbol means the sum of a number of quantities x, or whatever letter is within the bracket. If no letter is included within the bracket, and the whole symbol $S()$ is put at the bottom of a column of figures, it means the sum of the column. The sum of a number of such sums is denoted by $SS(x)$. So this symbol means the sum of a number of sums, themselves each of a quantity x. If the quantity is squared or cubed, its sum is denoted by $S(x^2)$ or $S(x^3)$, except that if the quantity within the bracket is a sum or difference of two quantities, it is denoted thus, $S(x+y)^2$. So this means the sum of a number of quantities $(x+y)^2$. If, on the other hand, the sum itself is squared, this is denoted by $S^2(x)$, or $S^2(x+y)$. So this means the sum of a number of quantities x, or $(x+y)$, the sum itself being squared.

In general, the summation symbol is used to mean the sum of all available observations, theoretically without limit if applicable, though it may be used to mean the sum of a limited number of observations. This is almost always clear from the context, and the insertion of limits is not usually really necessary, though it is often done. To do so, however, adds an appearance of complexity, and it is best avoided if possible. Sometimes a suffix is used, such as $S_a()$ or $S_f()$ for such things as the sum of an array or a family, terms to be described later. These suffixes will need careful attention from a student.

Another pair of symbols of which much use is made are n and N. The first, n, is used to denote the number of observations, while N is used to denote the number of 'degrees of freedom', an important term which is defined later. As a symbol for an indefinite number of integers, j is used, apart from its use in another sense in one of the examples. In this case no confusion is likely to arise.

The symbol of straight brackets, e.g. $|x-\bar{x}|$, is used to denote the positive value of the difference within the bracket. These differences are often squared in the analyses, and so in such cases the sign of the difference does not matter.

It must be explained also, at the start, that some of the terms of mathematical origin do not always bear quite the same shade of meaning when used in statistical methods as they do in ordinary mathematical usage. For example, a type of equation called a 'regression equation' may be derived statistically; although this looks like a mathematical equation, it cannot be treated in quite the same way. This is explained in the appropriate place.

For the benefit of students, some of the algebraic transformations used are derived in Chapter 16 at the end of the book. In that chapter they are numbered in sections. A reference is given, in the appropriate context, to the section of Chapter 16 containing the derivation. This saves interrupting the main argument. An understanding of the method of derivation should, however, be of help to students.

In all the analyses which follow, tabulations are necessary. They involve much work which is tedious, but this is unavoidable. Most of it is purely arithmetical, but some of the calculations have to be done very accurately, because they involve the differences of large quantities. Later stages can often be done on the slide rule. Careful setting out of the tabulations will be of help in making them seem less confusing, and foolscap paper ruled in quarter-inch squares is very useful. It is always well worth while to spend a little time and care in setting out and ruling up a tabulation.

CHAPTER 2

Presenting the Data: Counting; Tables; Diagrams

Most operations in statistical methods begin with a count of some operation or other, as a means of collecting the data. The actual method used in counting will naturally depend on the size and complexity of the problem being studied, and it is impossible to do more than give a few examples here. Of course, if there are comparatively few observations there is very little difficulty; we can count them directly.

Among the earliest examples used in this book are counts of timings of the speed of vehicles along stretches of road. These have been used to find out the effect of speed limits, or whether a speed limit is needed. These counts rarely timed more than about two hundred vehicles at a time, and the only information needed was the speed, apart from whether the vehicle was a private car, or a commercial vehicle which would already be subject to a speed limit. In these circumstances the counting of the speed values to make up the tables for analysis is straightforward. It is always, however, essential to be methodical. So that the counts can easily be done and checked, and also so that they can be easily intelligible to others besides the original observer, it is usual practice to mark each individual observation counted by a vertical stroke, thus |. This is done in the group or class in which the observation occurs, and then one marks each fifth observation by striking out the preceeding four with a diagonal stroke, thus 𝍸. A count of twelve then appears thus: 𝍸 𝍸 ||.

When the data are more complex, this procedure cannot be followed, and some form of card technique must be used. In the curve study already mentioned, the data were very complex, and this will serve as a useful example. The theory being studied was that drivers follow an invariable practice on all curves of using a fixed rate of change of

sideways—or 'lateral'—acceleration when entering a curve, except when racing. To test this theory, a recording accelerometer which records the force directly with a moving stylus on to a moving strip of paper was carried in a large number of vehicles, with different drivers, and on a large number of curves. It was clear at the start that if the theory should prove to be wrong—and very early in the study this was found to be the case—a number of different possibilities would open up, and a fresh theory would have to be found, if possible. It so happened that a suitable theory could not be found, but it still needed an extensive analysis to find that out. All this meant that as much data as possible had to be recorded for each individual curve. These included particulars of the road, the vehicle and the driver, the speed, and the hand of the curve, all of which were noted at the time of recording. The lateral force experienced on the curve, its rate of change, both at entrance and exit, and the proportion of the curve used as transition—i.e. when the steering wheel was being turned in entering and leaving the curve—were later calculated from the accelerometer record. Thus there were quite a number of observations to be recorded for each individual curve. The study was also made in war-time, and during the course of my ordinary duties, so that financial resources were limited.

In this case, the observations were all typed on to small cards identical in size, with each quality in the same place, and each denoted where applicable by a convenient letter or symbol available on the typewriter. If the observation was a number, that number would be typed in. To take an easy example, the hand of the curve was denoted by an R or an L in the chosen place on the card. When an analysis was to be made, the cards would be sorted into piles of the particular qualities needed and then counted, or sorted again into groups of other qualities, as necessary. An early analysis was to find out if the driver's behaviour varied with the hand of the curve; so the cards were sorted into two piles of L and R, and these two piles were then sorted again into the other qualities needed for the analysis. As a result of this operation it was found, as already mentioned above, to be improbable that the drivers' behaviour was affected by the hand of the curve, so that this could be neglected in the later analyses. At first sight this procedure of typing on to cards may appear to involve difficulties in the sorting, but a little practice soon makes it easy to spot the particular quality needed, in its place on the card.

A more elaborate method, saving much labour in sorting, but more expensive in equipment, is the use of some form of punched card. These are of two main types. In the simplest form the sorting and counting are done manually, in the other type these are done electrically. A form

of manually sorted card is the Paramount. This has holes punched round the outer edges of the card, and the top right-hand corner of the card is cut off so that it can be seen that all the cards are piled the right way when they are to be sorted. Each hole is taken to represent some quality or other, and when that quality is present on the card, the hole is opened out with a ticket punch. If that quality is required, the cards are piled, and a bodkin is passed through the hole representing it. Then when the pile is shaken, all the cards containing that quality will fall off the bodkin, while the rest will remain on it. The cards can be bought plain, or can be printed specially to requirements. The makers of the cards are usually prepared to give advice and help in their design for any particular operation.

The cost of the card, and its size, depend on the number of holes, and by careful design we can reduce this number, though possibly at

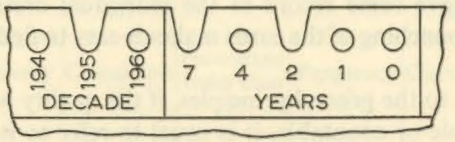


Fig. 1. Part of Paramount type of punched card, illustrating the design of holes for three decades.

the cost of some extra work in sorting. If there are only two categories of any one quality, we need allot a hole to only one of them. Thus for the hand of curve, which can only be left or right, it would only be necessary to use, say, one hole for right-hand curves. Then, in sorting, all the cards for right-hand curves would fall out, those for the left-hand curves remaining on the bodkin. A more elaborate example would be time, in years. Suppose we needed to card, for example, accidents over the years: it would need a very large and expensive card if we had one hole for each year, but we can represent a large number of years in a very few holes by alloting one hole to each decade, and then five other holes to the years in the decade. This is shown in Fig. 1, from which it will be seen that we can represent the thirty years between 1940 and 1969 in eight holes. In Fig. 1, the holes are shown punched for the year 1959. In sorting for this year, the cards would be piled, and the bodkin put through the hole marked 195. The cards falling out would be those for the 1950's. The rest would be put aside. The cards for the 195 decade would be piled again, and the bodkin put through the hole for 7. In this case, the cards falling would be for the three years 1957, 1958, and 1959. Then these would be piled again, and the bodkin put through

the hole marked 2. The cards which then fell out would be those for 1959.

The most labour-saving type of all are punched cards which can be analysed electrically. The relevant equipment is not always available, and it is more dependent on specialist cooperation; it is not described here. It may sometimes happen, however, that even when the electrical machines are available, cards of the Paramount type still have their uses. In Dorset, although the electrical machines are available, the Paramount type cards are being used for a study being made to find out the average life of surface dressings with different kinds of binder and of stone. For this type of study, which is a long-term one, the cards can act as the permanent office record of the dressings, as well as being used in the analyses. The study has been going on for thirteen years, but no strong conclusion has emerged yet because of the long life of some of the dressings, and also of the size of the sample needed. Thus it is of advantage to have some record of the individual dressings for office reference. The punching of the cards makes it easy to find any particular dressing.

Turning now to the general principles, if the quality we are considering is measurable or countable, it is usual to refer to it as the *variate*. The word 'stochastic' is sometimes used (adjectively), but has little to recommend it. We may think of the variate as the statistical equivalent of the algebraic variable, though as a special type with a probability attached to it. An algebraic variable will have an infinitely smooth variation from positive infinity to negative infinity, but our variate, though theoretically the same, will be found in practice to have a number of scattered values. We may therefore find that we have a number of observations of one value, but none of another, though in most of our work they will tend to be distributed about a central value. We will then want to know which is the more probable value, and we also want to present the scattered data in some clearer form.

Tables in general will be tolerably familiar, but there is one type much used in statistical methods, both as a means of presenting the data, and as a basis for calculations, which must be described in some detail. This is the *grouped table*, which is a very convenient way of presenting a large sample.

In this, the range of the variate is divided into a number of *groups* of an equal smaller range within the main range, and all observations within a group are counted together, as if they were all of the value of the centre reading of the group. Ten to a dozen is the most convenient number of groups, though the number is purely a matter of ease of calculation. It should preferably not be less than eight.

Some of the examples used in the book are counts of speeds of vehicles, and so speed may be used as an illustration of the process. The range of speeds observed was from about 12 miles/h to about 60 miles/h, and so twelve groups of 4 miles/h cover most of the observations conveniently, and the grouping used will be one of 4 miles/h. This range, of 4 miles/h, is called the *unit of grouping*.

Table 1 shows two sets of observations of the speeds of traffic, each set consisting of the sum of a number of separate timings taken at

TABLE 1. TIMINGS OF VEHICLE SPEEDS

(*a*) Within speed limits. (*b*) In lengths where a speed limit has been demanded

1	2	3	4	5	6	7
Speed V, miles/h	(*a*)			(*b*)		
	Frequency f	Cumulative f	Percentage faster than V	Frequency f	Cumulative f	Percentage faster than V
56/60				5	5	0·3
52/56	1	1	0·1	11	16	1·0
48/52	6	7	0·9	26	42	2·5
44/48	21	28	3·5	52	94	5·7
40/44	37	65	8·2	125	219	13·2
36/40	116	181	22·8	213	432	26·0
32/36	184	365	46·0	353	785	47·3
28/32	219	584	73·6	431	1216	73·2
24/28	141	725	91·4	299	1515	91·2
20/24	56	781	98·5	100	1615	97·3
16/20	11	792	99·9	35	1650	99·4
12/16	1	793	100·0	10	1660	100·0
Sum $S(f)$	793			1660		

different places and at different times. The first set, denoted by (*a*), covers timings taken on roads subject to the 30 miles/h speed limit. Some of these were in stretches fully built-up in all senses of the word, and some in stretches only partly built-up, but covered by street lighting, and so automatically subject to the limit. The second set, denoted by (*b*), were taken in stretches physically similar, but for various reasons—or absence of reason!—were not subject to the speed limit. In these cases a restriction had been demanded by some such body as a Parish Council.

These timings are made so that full and careful consideration can be

given to the need for a speed limit. There is much public demand for them, and great—though possibly misplaced—confidence is placed in them by many people. On the other hand they restrict motorists, and involve the punishment of infringers; so both justice and the public interest demand that they should not be imposed without very careful consideration, and that the case for them should be very strong.

The timings are here being considered as part of a general study of the speed limit, and the object of the present analysis is to find out if there is any evidence to show whether motorists behave differently in roads which are physically similar, but of which some have the limit, while some have not. In other words, to see whether the motorist adjusts his speed to an imposed limit, or to road conditions. Further analyses will be made with the same data later in the book.

In Table 1 the top line, containing numbers from 1 to 7, is merely the numbering of the columns for ease of explanation in the text. This practice is followed throughout most of the rest of the book. It does not form part of the analysis and is done merely for convenience in this book, and need not be done in the ordinary way.

Then column 1 sets out the range of speeds in groups of 4 miles/h. Thus the bottom line covers the range of speeds between 12 and 16 miles/h, the next line above those between 16 and 20 miles/h, and so on up the table. In practice, all speeds from 12 miles/h, up to, but not actually reaching, 16 miles/h, were included in the first group. Strictly speaking, any recordings of exact figures of 12, 16, 20 miles/h and so on should have been given half to the group above, and half to the group below. In a small sample this might be important, but with large ones such as these it is not worth doing.

The number of recordings in any one group is called the *frequency*, and is denoted by the letter f. The frequencies of the cars travelling on the set (a)—30 miles/h speed limit—stretches are given in column 2 of the table in the groups in which they occurred. Thus in these places, one car was timed at between 12 and 16 miles/h, eleven between 16 and 20 miles/h, fifty-six between 20 and 24 miles/h, and so on up the column. In column 5 the same thing has been done for the stretches on which the limit has been demanded. At the foot of the two columns they are added up to give the total number of cars timed in the two sets, or $S(f)$ according to the symbolism already explained. These totals will be seen to be 793 for the set (a) and 1660 for set (b). Both are large samples, but not abnormally large for traffic studies, in which the variations are apt to be wide. Columns 3, 4, 6 and 7 will be used later.

There are many types of diagram, most of which will be familiar.

Three types, which are of particular use statistically, will be described here. The first is the *histogram*, sometimes also called a *block diagram*. It sets out the quality being studied—speed in this instance—as abscissa,

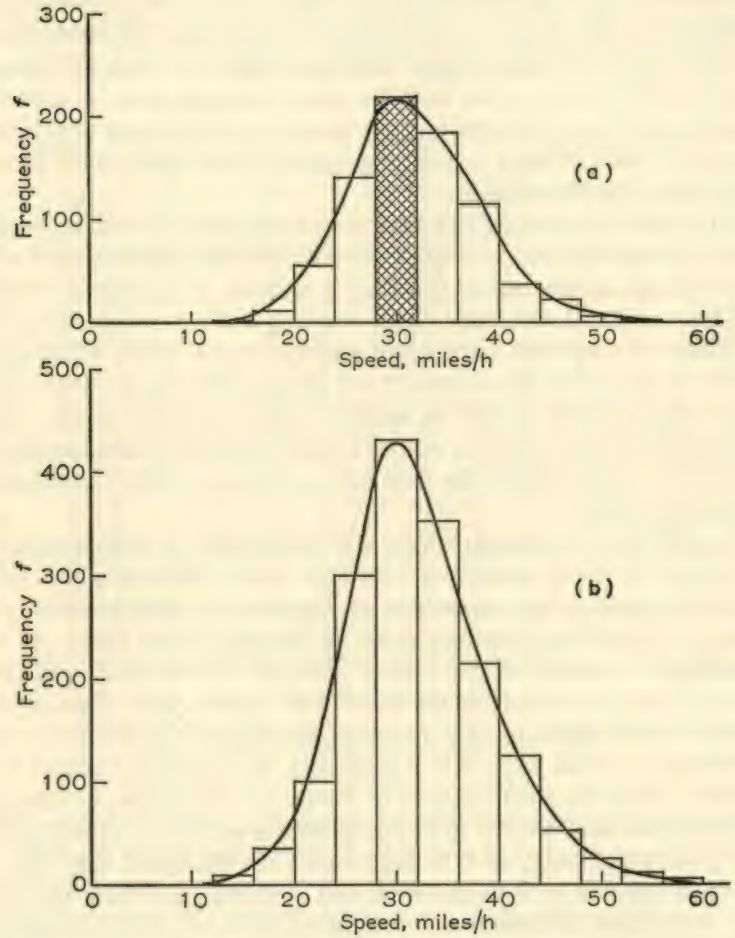


Fig. 2. Histograms of the speed timings in Table 1
(*a*) Timings within speed limits.
(*b*) Timings in lengths where speed limits had been demanded.

and the frequency as ordinate, showing the latter as blocks covering the corresponding groups to scale.

Histograms of the two sets of timings from Table 1 are given in Fig. 2. The speed is set out on the abscissa, and also the groups corresponding

to those on the table, to scale. Then on each group is drawn a rectangle of height according to the frequency of the recordings in the group, measured to scale on the ordinate. Thus, in the 28/32 miles/h group of the set (a) timings, the frequency from the table, column 2, is 219, so that the height of the rectangle to scale on that group is 219. That rectangle or block is shown cross-hatched on the figure. The same thing is done for all the other groups. The figure shows the two histograms which result. It also shows how the whole diagram gives us a useful impression of the distribution of the speeds in the two sets of timings, and that it does so more realistically, and in a more easily understandable form, than the table.

If the unit of grouping had been very small, and the sample a very large one indeed, a nearly smooth curve would result, getting more and more smooth as the unit of grouping is reduced. If an infinite sample had been obtained, the shape of the resulting curve would probably be as shown in a full line on the two histograms, on which it has been drawn by inspection. From this, it will be seen that the distribution of the timings covered a sort of roughly bell-shaped and nearly symmetrical distribution about a central value, though with some tendency for it to be skew towards the slow side, and with a slight tail towards the higher speeds.

Another type of diagram which is of importance in traffic studies is the *percentile curve*, sometimes called an *ogive*. When this has been plotted it enables the percentage of observations falling below—or above, at choice—any desired value, to be read off the curve. As an illustration of the use of this type of curve, it may be mentioned that in the U.S.A. it is becoming the practice to impose speed limits of the nearest round figure to that speed which 85% of the traffic is not exceeding, or which 15% of it is exceeding. The argument behind this practice is that the great majority of drivers are reasonable, competent people who can be trusted to decide what is a safe speed to use under any given conditions; so if a limit based on the speed they use is imposed, some at least of the other sort of driver may think that the limit is justified, and observe it themselves. It also has the advantage that it brings an element of reason into the limit which is absent in the British fixed one of 30 miles/h, especially as this has been imposed in many cases quite arbitrarily. It is claimed that this 'percentile limit' has been successful in reducing accidents, especially fatal accidents. Its adoption in this country has been suggested, and in Dorset a percentile curve is always plotted for the timings made in a place where a speed limit has been demanded, for the guidance of the committee considering the matter.

The method of computing a percentile curve is included in Table 1 (page 9), in columns 3 and 4 for the set (a) timings and in columns 6 and 7 for the set (b) timings. The first step is to sum the frequencies cumulatively. This can be done indifferently from the top or the bottom of the column, entirely at the choice of the computer. For reasons which will appear later, it is here done from the top. This operation is done in columns 3 and 6. Taking column 3, for the set (a) timings, the first

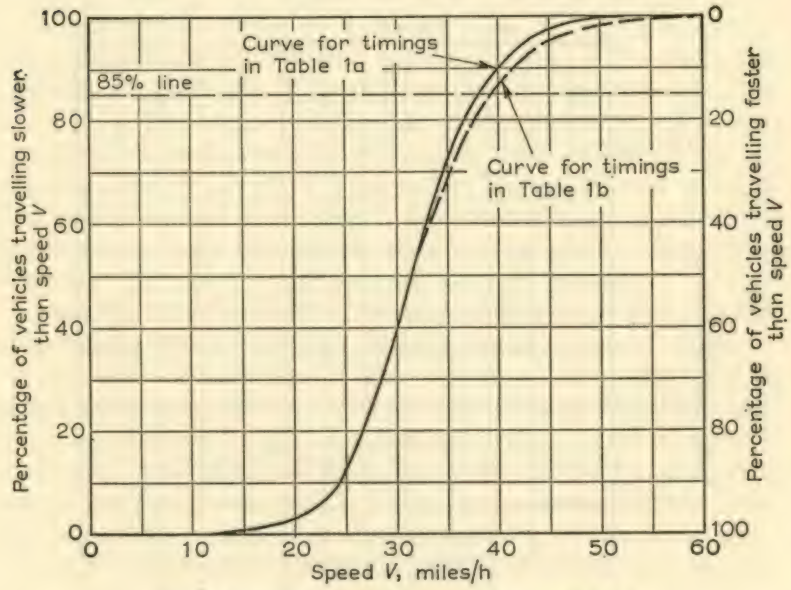


Fig. 3. Percentile curves for the timings in Table 1

frequency is 1, in the 52/56 group, and this is entered in the same line in column 3. In the next line the frequency is 6, and this figure is added to the first one, of 1, to give 7 in the next line of column 3. The next frequency below, 21, is then added to this, giving 28, and entered into the next line of column 3, and so on down the column. The last figure, in the last line in which a frequency occurs—here the 12/16 group, should give the sum of the total number of timings, i.e. $S(f)$, if the working is correct. A similar process is done for the set (b) timings in column 6.

The next step is to work out, and enter into column 4—or 7 as the case may be—the percentage of the total number of observations represented by the entry in each line of column 3 or 6. Thus, again

taking column 3, the number in the first line, namely 1, is 0·1% of the total of 793, and this figure is entered into the same line of column 4. The next number in column 3, namely 7, is 0·9% of 793, and so on down the column.

In drawing the diagram (Fig. 3), the speeds are again plotted as abscissa, and the percentage as ordinate. On one side of the diagram, on the ordinate line, the percentages are plotted upwards from the bottom. This is done on the left side in Fig. 3, and the line shows the

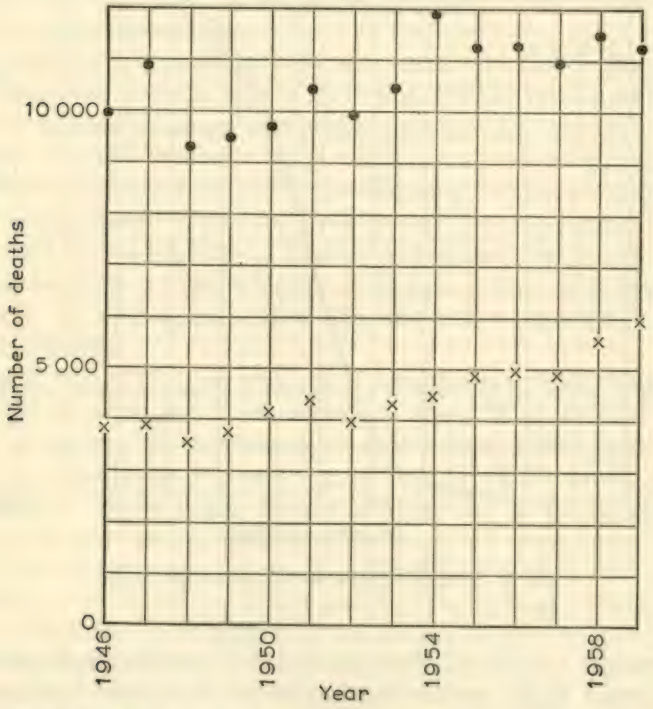


Fig. 4. Dot diagram: accidental deaths in England and Wales from 1946 to 1959

● Deaths in accidents other than motor vehicle traffic accidents.
× Deaths in motor vehicle traffic accidents.

percentage of vehicles travelling slower than the speed V, given by the intercept of the plotted curve on the abscissa. On the other side the percentages are plotted downwards, and this line reads the percentage travelling faster than the speed V.

In this case, the curve is plotted from the top, as the calculation has

been done that way. The percentage figure in column 4 or 6 is plotted against the lower speed of its group range, because the percentage is that of speeds greater than the lower speed of the group. For example, the percentage in the 40/44 group of set (a) is 8·2, and this is the percentage of cars travelling at over 40 miles/h. If the percentage had been calculated from the bottom, the reverse would apply, and the percentages would be plotted against the upper speed for the group. Here, for the set (a) curve, 0·1% on the right-hand line is plotted against 52 miles/h, 0·9% against 48 miles/h, and so on. The points so plotted are then joined by a smooth curve. In Fig. 3 the set (a) timings are plotted in a full line, and those for the set (b) timings in a dashed line. Below the 60%–40% line the two curves coincide. The 85%–15% or 85 percentile line has been drawn on, and it will be seen that the 85 percentile speed for curve (a) is about 38 miles/h, and for curve (b) about 39 miles/h. The curves also indicate that the principal effect of the imposed limit seems to have been slightly to reduce the highest speeds, with little or no effect on the speeds below about 33 miles/h. The reduction is, however, not very marked.

A percentile curve is usually of the shape shown on the figure. Different conditions will affect its scale, but not usually its general shape.

If the data contain values of two variates x and y, they may be plotted, using a dot—with or without some distinguishing mark such as a cross, circle or triangle—for each observation, in the usual way. In our work, the dots will usually be scattered, sometimes in an approximation to a line of some sort, as in Fig. 4, but sometimes quite indiscriminately. This is called a *dot diagram*, and Fig. 4 is an example.

Plotting a dot diagram sometimes helps to give some idea whether there is likely to be some relationship between x and y. A quick approximate method of finding this out is given by Quenouille.[2] The diagram plotted in Fig. 4 will be used in an analysis in Chapter 10, which will show that there is a probable relationship in both cases, as one might expect from examining the figure.

Example 2.1. The figures on page 16 show the winter rainfall for the Thames Basin, taken from a paper by Andrews.*

Arrange these data into a grouped table, draw a histogram and a percentile curve. From the latter curve deduce the percentage of water years with rainfall above 22 in per winter, and below 12 in per winter.

Note: To show the distribution, it is suggested that the unit of grouping should be 1 in of rain.

* Andrews, F. M.: 'Some Aspects of the Hydrology of the Thames Basin', *Proc. Instn. Civil Engrs.* **21** (1962), Paper No. 6568.

Water year	Winter rainfall, in	Water year	Winter rainfall, in	Water year	Winter rainfall, in	Water year	Winter rainfall, in
1883–84	12·8	1903–4	20·1	1923–24	15·5	1943–44	9·5
1884–85	12·9	1904–5	10·9	1924–25	18·5	1944–45	15·7
1885–86	15·3	1905–6	14·6	1925–26	15·2	1945–46	13·8
1886–87	16·6	1906–7	14·9	1926–27	18·5	1946–47	18·9
1887–88	12·8	1907–8	18·3	1927–28	17·6	1947–48	12·2
1888–89	13·9	1908–9	10·5	1928–29	12·8	1948–49	12·4
1889–90	11·6	1909–10	16·7	1929–30	23·8	1949–50	18·3
1890–91	8·4	1910–11	17·5	1930–31	13·2	1950–51	20·7
1891–92	17·3	1911–12	23·9	1931–32	9·8	1951–52	16·4
1892–93	13·0	1912–13	17·5	1932–33	16·2	1952–53	13·9
1893–94	15·1	1913–14	17·3	1933–34	8·5	1953–54	12·3
1894–95	16·3	1914–15	23·1	1934–35	14·7	1954–55	17·0
1895–96	14·8	1915–16	22·6	1935–36	21·3	1955–56	13·3
1896–97	17·3	1916–17	17·2	1936–37	21·2	1956–57	14·0
1897–98	9·0	1917–18	12·4	1937–38	12·1	1957–58	14·9
1898–99	16·5	1918–19	19·2	1938–39	19·0	1958–59	15·7
1899–00	17·1	1919–20	14·2	1939–40	19·8		
1900–1	13·8	1920–21	10·9	1940–41	21·4		
1901–2	11·0	1921–22	14·0	1941–42	12·5		
1902–3	14·0	1922–23	15·2	1942–43	16·5		

My thanks are due to the author of the paper, and to the Institution of Civil Engineers, for the use of these data.

279.5 331 328.1 239 Total 76 1177.6

CHAPTER 3

The Population and its Distribution

The subject under study is called the *population* or sometimes the *universe*. This may be thought of as the aggregate of an infinite number of observations of the type being considered, though a population may be divided into sub-populations for some purposes. Thus we may consider the population—in the conventional sense of the word—of the whole world, that is, man as a whole. But we may also consider the population of any one country or region of the world, and it would be possible, and legitimate, to compare the characteristics of the sub-population with those of the whole population, or with those of other sub-populations. While under analysis, these sub-populations would themselves be populations in the statistical sense of the word.

In such cases, while the word 'infinite' may be used to describe the population, it is not necessarily used in its full mathematical sense. It can be used in this sense, but statistically it may also be used to denote a large number which may yet be countable.

In general, if we are considering measurements, the population would be an infinite number—in the mathematical sense—of the type of measurement being studied, and these would constitute the variate. But the data we actually have are inevitably only a part of the population, and this part is called the *sample*. The problem may be to find out the characteristics of the population from those of the sample, or, if two or more samples have been gathered, we may want to know whether they can be thought of as belonging to the same population, or as drawn from different ones. But, before we can consider how this can be done, we must study how the population is distributed, because if it is not distributed in some way it will have no scatter, and can be considered as a single number.

DISTRIBUTIONS

(i) *The Binominal Distribution*

The simplest and most familiar case of a chance distribution is the spin of a coin. If the coin is true, the probability of the fall of a head is the same as that of a tail. Probabilities are always expressed in fractions or decimals, and their sum in any given case is always unity; so we express the probability of either the head or the tail of a coin falling as $\frac{1}{2}$, and the total probability is $(\frac{1}{2}+\frac{1}{2})=1$. We might alternatively say that the chance is one in two of either event.

This does not mean, of course, that we ought to expect that every time a head falls the next event should be a tail. We know that it is not. We merely mean that in a very large number of throws of a true coin, the proportion of heads or tails tends to be roughly equal, and that the larger the number of throws, the more will the two proportions tend to be equal. But, however many throws have been made, and whatever the result has been hitherto, the probability of a head in the next throw is still $\frac{1}{2}$.

In general terms, it is usual to call the occurrence of some event, at choice, a success, and to denote its probability by p. Then its non-occurrence—or the occurrence of any other event—would be called a failure, and its probability would be denoted by q. If for the spin of the coin we call the fall of a head a success, we have that

$$(p+q)=(\tfrac{1}{2}+\tfrac{1}{2})=1$$

There may, of course, be more than two probabilities if there are more than two different events which could take place. An example will be given below. But even so, the probabilities will still add up to unity.

We may now go on to an equally familiar, but more complicated, case, that of the ordinary six-sided die. If we select the fall of a six as a success, and denote its probability by p as usual, and call the fall of any other number a failure, with its probability as q, we would then have that $p=\frac{1}{6}$, since there is one chance in six that a six will fall. Then $q=\frac{5}{6}$, and

$$(p+q)=(\tfrac{1}{6}+\tfrac{5}{6})=1$$

Extending all this, and putting it into more general terms, it was shown by the French mathematician Bernouilli in the eighteenth century that the probability of getting a given proportion of either successes or failures in a random sample of 1, 2, 3, 4,... r trials will be given by the terms of the expansion, by means of the binomial theorem, of the expression

$$(p+q)^r$$

For this reason, the distribution is known as the *binomial distribution.*

Returning to the dice, if we throw two of them at a time, there are two alternatives, i.e. a success or a failure—so that in this case $r=2$, and by the expression given, the probabilities in a single throw are

$$(p+q)^2=(\tfrac{1}{6}+\tfrac{5}{6})^2=\tfrac{1}{36}(1+10+25)$$

If, then, we threw the two dice 36 times, so as to get whole numbers,

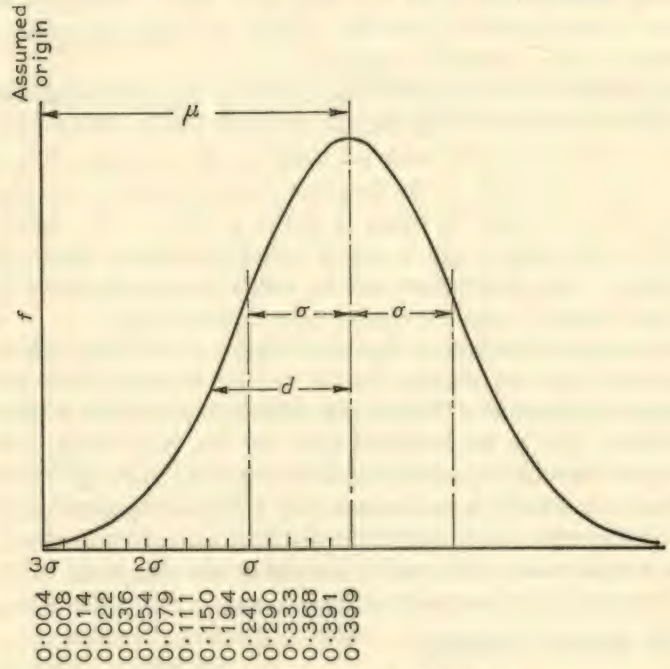


Fig. 5. The normal distribution

The figures are the ordinates to the curve at intervals of $0{\cdot}2\sigma$ measured from the centre-line of the curve and calculated with $\sigma=1$.

we would expect to get two sixes—i.e. two successes—once; one six ten times; and no sixes at all twenty-five times.

(ii) *The Normal Distribution*

The binomial distribution is unsymmetrical and discontinuous, but it becomes more and more continuous and symmetrical as r increases. When r increases without limit, the distribution becomes completely symmetrical and continuous; one can show that it is expressed by the equation

$$f=\frac{1}{\sigma\sqrt{(2\pi)}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \ldots\ldots \quad (1)$$

In this expression, x is the variate, f is the probability density of any value of x, e has its usual meaning of the base of Napierian logarithms, and μ and σ are functions which will be described later. They have replaced the probability p and the power r which, it will be seen, do not form part of the equation.

A *probability density* is a function which has the property that when it is plotted in a curve such as is shown in Fig. 5 (page 19), the area of the curve between two values of x such as x_1 and x_2 multiplied by the number of observations gives the number of times the values of x between x_1 and x_2 would be expected to occur.

The distribution represented by equation 1 is the most important one with which we have to deal, and the analyses which will be described in most of the rest of this book are based on its properties. It is called the *normal distribution*, or the *Gaussian distribution*, after the German mathematician Gauss. Its shape is shown in Fig. 5. The figure also shows the functions μ and σ, which will be discussed below. About two-thirds of the observations will lie within the area bounded by the two lines distant σ from the central value denoted by μ.

The binomial distribution, and the Poisson distribution, which will be described next, are discrete. That is, we have in them a finite number of different probabilities. The normal distribution, and the exponential distribution, also to be described later, on the other hand, follow a continuous curve. The probability of the occurrence of any deviation d from the central value μ, as shown in Fig. 5, is given by the area cut off, outwards from the centre, by the vertical lines located by the dimension d. If d is equal to σ, then about one-third of the area is cut off by the two vertical lines located by the dimension σ, as mentioned above.

(iii) *The Poisson Distribution*

It has been said above that the probabilities given by the binomial distribution are found by the expansion of the expression $(p+q)^r$, but when either p or q in this expression is of the order $1/r$, the probabilities of getting $0,1,2,3,4,\ldots j$ events are given by the terms of the series

$$e^{-m}\left(1,\ m,\ \frac{m^2}{2!},\ \frac{m^3}{3!},\ \frac{m^4}{4!},\ldots\ \frac{m^j}{j!},\ldots\right)$$

This is called the *Poisson series*, after the French mathematician Poisson. It applies when the probability of an event occurring is very small, but when the number of possibilities of its occurring is very large. This series occurs in traffic studies, and will be considered in detail below.

(iv) *The Exponential Distribution*

When events are randomly distributed in time, their distribution may

be covered by what is known as the *exponential distribution*. This is defined by the expression

$$f_{(t)} = \lambda e^{-\lambda t} \quad \ldots\ldots\ldots \quad (2)$$

in which λ is a number derived from the data, such as the rate of flow of traffic, $f_{(t)}$ is a probability density, and e has its usual meaning of the base of Napierian logarithms. This distribution will also be described in detail below.

(v) *Other Distributions*

There are many other types of distribution, and diagrams of a few of them are given in Fig. 6. Some of them are not uncommon. The type

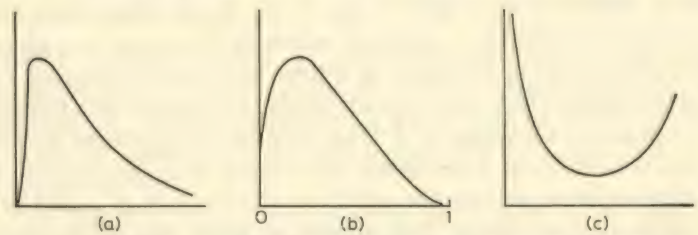


Fig. 6. Some types of distribution other than the normal distribution

shown in Fig. 6*b* was found for the proportion of the total length of a curve used by drivers as transition, in the curve study already described. This proportion is a ratio, and so by definition varies from zero to unity. The curve of Fig. 6*c* represents roughly the observed rate of certain types of accident when plotted against the age of the victims. That is to say, the accident rate is higher in the very young or the very old, with a minimum in middle age.

Changing the Distribution

It is possible, and legitimate, to change the distribution of a quality into another distribution by changing the function in which it is expressed. This is sometimes important, because some of the tests described below depend on the quality studied being distributed normally, or nearly normally. If its distribution is markedly not normal it can often be made more normal by using some function of it such as the logarithm or the square. This operation may sometimes also be used to save work in the analysis—an aspect of the matter to be discussed in a later chapter.

Sometimes it is not even possible for the quality to be normally distributed. An example of this, mentioned in the preceding Section v

of this chapter, was the proportion of transition, plotted in Fig. 6*b*. The same remark would apply to any proportion, since this must, by definition, vary from unity to zero. The normal distribution must in theory vary from positive to negative infinity. The fact that extreme values are rare does not alter the principle.

This can be avoided by expressing a proportion in the form of a function z, defined by the expression

$$z = \tfrac{1}{2}[\log_e P - \log_e(1-P)] \quad . \ . \ . \ . \quad (3)$$

In this expression, P is the proportion, varying from 0 to 1, and the logarithm is the natural logarithm.

This looks complicated, but offers no difficulty if cards are being sorted as described in Chapter 2 (pp. 5 *ff.*). Since proportions sometimes occur in practice, a table of values of P against z is given in Table A at the end of the book, in a form suitable for grouping z. All that is necessary, then, is to sort the cards into groups of z instead of P. For example, all values of P from 0·74 to 0·82 inclusive would be put into the +0·75 to +0·50 group of z, values of P from 0·37 to 0·27 inclusive would be put into the −0·25 to −0·50 group of z, and so on. But values of P of exactly 0·50 should be given half to the positive group of z above, and half to the negative group of z below zero. A similar type of procedure could be followed for any other transformation.

CHAPTER 4

Parameters as Measures of Location

Certain functions known as *parameters* are used to epitomize the characteristics of a given population. In a fully normal distribution, for example, the two parameters μ and σ give us all the information we need to know about the population. The next stage is to consider these parameters.

The first main parameter measures the location of the distribution, and is, roughly speaking, the central value. It is often loosely called the 'average', which word is more generalized in statistical usage than in popular idiom. There are several types of this, and the three principal are:

(a) The *mean*. This is what most people think of when they refer to the average. It is the most important of all the averages. It is defined as the sum of all the observations, divided by their number.

(b) The *median*. This is the central value, that is to say, the value which has an equal number of observations larger and smaller than itself.

(c) The *mode*. This is the observation which occurs most frequently.

These averages have different uses, but the mean is the most important because it is the only one which can conveniently be dealt with mathematically. An illustration of the misuse of averages is given by Huff.[3] He describes some uses—or abuses!—made of the expression 'average income' of a community in contexts in which the type of average is carefully left undefined. The distribution of incomes is almost certain to be of the type shown in Fig. 6*a*. There are a large number of people in the smaller income groups, and a very small number in the very large income groups. But when the incomes are added up to find the mean income, the few large ones contribute a disproportionate amount

to the total, and so the mean works out too large to give a genuine idea of the community income average—which is probably the effect intended. The median, or possibly the mode, would give a much better idea.

In the normal distribution the mean, the median and the mode are all the same, and they are the dimension denoted by μ, in Fig. 5 (p. 19).

At this stage I must digress slightly to mention that it is an established convention (and a very important and useful one) to use Greek letters to denote the parameters of the population—hence the use of μ and σ in Fig. 5. So the letter μ is used as a generalized symbol for the mean of a population. On the other hand, the parameters derived from the sample are only estimates of those for the population, and so it is usual to distinguish them by denoting them by means of ordinary letters.

We have a difficulty with the mean, however, in that there is a variety of symbols for the variates we use, and we may also have several variates at once in the same problem. Thus we clearly cannot use one symbol for the mean, and the usual practice is to put a bar over the symbol used for the variate: $\bar{x}$, $\bar{y}$, $\overline{V}$, etc. These symbols may denote either the mean of the population or its estimate from the sample. This is unavoidable, but should not lead to any confusion, because it should be unambiguous from the context which is meant. If not, it may be necessary to make the matter clear. If a further generalized symbol for the mean is needed, and instances of this occur below, the letter m is used for the estimate from the sample.

Then, if we denote the sum of a number of observations of a variate x by $S(x)$, and the number of observations by n, the mean is, by definition, given by the equation

$$\bar{x} = \frac{S(x)}{n} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

From this it follows (see Section 16.1 in Chapter 16) that

$$S(x-\bar{x}) = 0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4a)$$

The term $(x-\bar{x})$ is called a *deviation from the mean*, or sometimes a *deviation*, and this term is much used in the rest of the book.

For small samples the mean can be calculated easily and directly by adding up all the observations and dividing the resulting sum by the number of observations. This is done in Table 2, which shows two small samples of fuel consumption in two different types of motor lorry. One type is driven by ordinary motor spirit and is of 3-ton nominal capacity, and the other type is 5-ton diesel-driven. All the lorries in each sample are of the same make and type. The working should be obvious from the table.

With such small samples this is very simple, but as the size of sample increases it becomes more and more laborious, and a shortened method must be used. This method is, by extension, used to calculate the next parameter, and so it will be described in some detail, using for the purpose the second of the two sets of timings, set (b), previously used in Table 1 (page 9). This is carried to Table 3.

TABLE 2. CALCULATION OF THE MEAN: FUEL CONSUMPTION OF LORRIES

1	2
Fuel consumption, miles/gal	
3-ton petrol	5-ton diesel
7·98 8·65 9·08 10·62 10·52 10·61 10·14	14·94 15·83 16·71 18·25 15·42 16·29 15·94
$S(\ )$ 67·60	113·38
Mean $\frac{67{\cdot}60}{7} = 9{\cdot}66$	$\frac{113{\cdot}38}{7} = 16{\cdot}20$

In Table 3 the speeds are, as before, given in column 1, and the frequencies are entered into column 3. But if we used these columns directly, multiplying the frequency by the central speed of the group, as would be possible, we should get very large figures, and this would make the whole process cumbersome. For instance, we have 213 cases of speeds between 36 and 40 miles/h. This speed averages at 38 miles/h, and so the total of speeds in that group would be $213 \times 38 = 8094$. In the later analyses the figures would become even more cumbersome.

We can get over this difficulty without any loss of accuracy by using a *working unit* of the range covered by one group. This is always worth doing. In this case, then, we would use a working unit of four miles per hour. We then number these working units from a *working mean*, or *working zero* as it is sometimes called, and we denote them by x. The working mean can be chosen anywhere in the table which is found convenient. In many books on the subject it is recommended that the working mean should be taken in the group in which the true mean

appears to lie. Thus in Table 3, if this method were used, it would probably be taken in the 28/32 or 32/36 group, i.e. it would be made either 30 or 34 miles/h. In this method we would then set out x in column 2 of the table from the chosen zero, upwards as positive and downwards as negative. This involves signs, which is a drawback, and it is not convenient for the alternative summation method to be described below.

TABLE 3. CALCULATION OF THE MEAN: VEHICLE SPEEDS

1	2	3	4	4a
Speed V, miles/h	x	f (from column 5 of Table 1)	fx	Alternative summation method C1
56/60	11	5	55	5
52/56	10	11	110	16
48/52	9	26	234	42
44/48	8	52	416	94
40/44	7	125	875	219
36/40	6	213	1278	432
32/36	5	353	1765	785
28/32	4	431	1724	1216
24/28	3	299	897	1515
20/24	2	100	200	1615
16/20	1	35	35	1650
12/16	0	10	0	—
		1660 $= n = S(f)$	7589 $= S(fx)$	7589 $= S(fx)$

$$\bar{x} = \frac{S(x)}{n} = \frac{S(fx)}{n} = \frac{7589}{1660} = 4{\cdot}5716 \text{ in working units}$$

$$\bar{V} = 4 \times 4{\cdot}5716 + 14{\cdot}00 = 18{\cdot}29 + 14{\cdot}00 = 32{\cdot}29 \text{ miles/h}$$

The method followed here is to take the working mean in the lowest group and then number x upwards from this, as has been done in column 2 of Table 3. This table has deliberately been chosen from a very large sample; yet it can be seen that the method does not involve inconveniently large numbers. There is really very little to be said for the method mentioned above.

There are now two ways of working out $S(x)$, the quantity we need for finding $\bar{x}$ (the mean in working units) and so for deriving $\bar{V}$ (the mean speed in miles per hour). The first is the more usual method. This is to multiply the value of x in any line by that of f in the same line, entering the result in the fx column, column 4. Thus, taking the

40/44 group, in which 125 vehicles were counted, and for which $x = 7$, $fx = 7 \times 125 = 875$. The rest of the lines would be done similarly. Then the sum of the fx column is $S(fx)$, which is $S(V)$ in working units. This sum is seen to be 7589.

There is another method which only involves addition and is very quickly and easily done. It also provides a saving in work in the later analysis. It is shown by column 4*a* of Table 3, which is headed C1, because it uses the cumulative summation method described for Table 1 (page 9), with the difference that the line for the working zero is not included. This method, however, would probably be less likely to lead to slips in the computation if the working zero were taken in the group below the bottom line containing a frequency. In Table 3, then, it would be taken one group below the present zero. It has not in fact been done here so as to make the two methods consistent. One important proviso must be made. If some of the lower groups, above the working zero, contained no frequencies, the summation would still have to be carried down to the line above the working zero. Thus supposing column 3 of Table 3 contained no frequency in line 2 from the bottom, when $x = 1$, instead of 35, the figure of 1615 would be repeated in column 42 from the line above, instead of 1650. The reason for this can be seen from the examination of the derivation of the method in Section 16.3 in Chapter 16.

Column 4*a* of Table 3, then, contains the cumulative summation, from the top, of the f values in column 3, omitting the bottom line. Its last term should equal the number in the sample, less the frequency in the line containing the working zero, if the summation has been done correctly. Here the last term is 1650, which is 10 less than n, the number of observations in the sample, equal to 1660. The sum of this column will be seen to be $S(fx)$. The method provides some saving in work, and this saving will be greater in the later analyses. If it was used, column 4 would not be needed, and it would not be really essential to include column 2, the x column, provided it is remembered where the working zero is.

Then, whichever method has been followed—and the matter is entirely one of personal choice—$\bar{x}$ and $\bar{V}$ are worked out below the table. The mean speed in working units, $\bar{x}$, is $S(fx)$, the sum of the speeds in working units, divided by n, the number of vehicles timed. This gives 4·5716. Then, to find $\bar{V}$, this figure is multiplied by the working unit of four miles per hour, and added to the value of the working mean, which is 14 miles/h, giving a mean speed of 32·29 miles/h.

If we put this operation into symbols, expressing the quality being

studied—in this case speed—by X, the working unit by U and the working mean by $\bar{w}$, the method is illustrated by the equation

$$X = U\bar{x} + \bar{w} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

Example 4.1. The following figures* give the road accident death rate per hundred million vehicle-miles in two states of the U.S.A. over a period of 13 years. Calculate the mean death rate for the two states.

Year	1947	1948	1949	1950	1951	1952	1953
Connecticut	4·4	4·6	3·4	4·0	3·9	3·2	3·6
Rhode Island	4·8	2·9	3·0	3·8	3·0	3·0	2·8

Year	1954	1955	1956	1957	1958	1959
Connecticut	3·1	3·8	3·2	3·0	2·6	2·5
Rhode Island	2·4	3·0	2·4	2·7	2·5	3·0

* Kindly supplied by Dr. Daniel P. Moynihan, Syracuse University, N.Y., U.S.A.

Example 4.2. Using the data of Example 2.1, calculate the mean annual winter rainfall in the Thames basin for the years 1883 to 1959.

CHAPTER 5

The Standard Deviation and its Estimate from the Sample

If we only know the mean of a population we have not enough information about it. We must for completeness know something about its spread, or scatter, about the mean. Let us suppose that we have a mean of 100, and that we want to know whether we could expect another observation of 95 to be part of the same population. So far, we cannot say anything about it, but let us go on to suppose that we had two more original observations, from which we had derived the mean, of 99 and 101. We might at first sight say that 95 was exceptional; but, after all, we have only three readings, and so we cannot say with complete confidence that 95 was really exceptional, though the case would be different if we had a larger number of observations between 99 and 101. On the other hand, if the three readings were 90, 100 and 110, we could then say with confidence that another of 95 would not be unexpected. Thus we need some measure of the scatter about the mean before we can really know much about the population, and this measure must also be related to the number of observations.

For this purpose, we use a parameter called the *standard deviation.* This is denoted by σ for the population. It is the dimension shown in Fig. 5 for the normal distribution, and is the function denoted by the same symbol in eqn. 1. The standard deviation is the square root of a function known as the *variance*, which is defined by the equation

$$\sigma^2 = \frac{S(x-\mu)^2}{n}$$

The numerator in this equation, when expressed for a sample in the form $S(x-\bar{x})^2$, is one of the most important quantities used in statistical methods, and it is called the *sum of squares.* The words 'of deviations

from the mean' are understood as being added to the term. The variance, then, is the sum of all the squared deviations from the mean—the sum of squares—divided by the number of observations in the population.

For any finite number of observations, the sum of squares can obviously be calculated by working out the deviations, squaring them, and adding. But while this is possible for a small number of observations, it clearly becomes impossibly laborious for a large number of observations. It can, however, be derived in a form which is much easier to calculate (see Section 16.2 of Chapter 16):

$$S(x-\bar{x})^2 = S(x^2)-\bar{x}\,.\,S(x) \quad \text{or}$$

$$= S(x^2)-\frac{S^2(x)}{n} \qquad \ldots\ldots \quad (6)$$

Either of the two forms of this expression can be used according to arithmetical convenience. The second form may sometimes be easier to calculate accurately. The sum of squares is always positive.

The equation given above for the variance σ^2 has not been numbered because it can never be used in this form. The reason for this, which introduces another very important conception, is that we are dealing with an infinite population, at least in theory; so for practical reasons we cannot calculate the variance from that equation, because both terms would be infinite, and so would be too large to count. We also do not know the value of μ, only that of m, the mean of the sample. Thus in practice we have to estimate the variance of the population from that of a sample. But it is most unlikely that the variance of the sample will be the same as that of the population, which, in general, it tends to underestimate. It is thus most important to distinguish the two functions. For this reason the quantity for the sample which corresponds to the variance of the population is called the *mean square*. Its square root, which corresponds to the standard deviation of the population, is called the *root mean square*, and it is denoted by s, on the principle already mentioned (page 24) of using ordinary letters for the estimates of parameters made from a sample.

These definitions are of extreme importance, and must be fully understood. They can be summed up in a small table, as shown in Table 4.

Now let us go on to consider the smallest possible sample, one consisting of two observations, denoted by a and b. Their mean is obviously $\frac{1}{2}(a+b)$, and by using the second form of eqn. 6 we derive their sum of squares as

$$a^2+b^2-\tfrac{1}{2}(a+b)^2 = \tfrac{1}{2}(2a^2+2b^2-a^2-2ab-b^2)$$
$$= \tfrac{1}{2}(a^2+b^2-2ab) = \tfrac{1}{2}(a-b)^2$$

This is clearly based on one comparison, that between a and b.

Next we add another observation to the sample, making it one of

TABLE 4

Parameter of the population		Estimate from a sample	
Name	Symbol	Name	Symbol
Standard deviation	σ	Root mean square	s
Variance	σ^2	Mean square	s^2

three observations a, b and c. The mean is now $\frac{1}{3}(a+b+c)$. Again using the second form of eqn. 6, we derive the sum of squares as

$$a^2+b^2+c^2-\tfrac{1}{3}(a+b+c)^2 = \tfrac{2}{3}(a^2+b^2+c^2-ab-ac-bc)$$

To this term we first add, and then subtract, the sum of squares from the last sample we have taken. The sum of squares for the sample of three observations then becomes

$$\tfrac{1}{2}(a-b)^2+\tfrac{2}{3}(a^2+b^2+c^2-ab-ac-bc)-\tfrac{1}{2}(a-b)^2$$
$$= \tfrac{1}{2}(a-b)^2+\tfrac{1}{6}\{4a^2+4b^2+4c^2-4ab-4ac-4bc-3a^2+6ab-3b^2\}$$
$$= \tfrac{1}{2}(a-b)^2+\tfrac{1}{6}\{a^2+b^2+4c^2+2ab-4ac-4bc\}$$
$$= \tfrac{1}{2}(a-b)^2+\tfrac{1}{6}\{a+b-2c\}^2$$
$$= \tfrac{1}{2}(a-b)^2+\tfrac{2}{3}\{\tfrac{1}{2}(a+b)-c\}^2$$

The first term of this expression is the sum of squares we have already derived for the sample of two observations, and the second is a single comparison between the mean of the first two observations and the third one. So the number of comparisons available for finding the sum of squares of the sample of three observations is two.

There is, of course, no special reason why we should make the comparisons between a and b taken together and then compare them with c. We could equally have chosen to compare a and c with b, or b and c with a, without affecting the issue, but, however we do it, the result is two comparisons.

If we repeated the process with four observations we would, as the result of a more elaborate algebraic process, arrive at a comparable result, that we would have three independent comparisons that we can make in the data to arrive at the sum of squares.

Carried still further, we would arrive at the conclusion that if we have a sample of n observations, we always have $(n-1)$ independent comparisons in the data for computing the mean square since we

have used one up in computing the mean, so to estimate the mean square we must divide the sum of squares by this number. This gives us a general expression for the mean square of the sample,

$$s^2 = \frac{S(x-\bar{x})^2}{(n-1)} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

The denominator of this expression is called the number of *degrees of freedom*. A degree of freedom is defined by Mather[4] as 'a comparison between the data, independent of the other comparisons used in the analysis'. In general, having calculated a parameter from the data, we have lost a degree of freedom. Thus when calculating the sum of squares, we must first calculate the mean, which costs us one degree of freedom; so for arriving at the mean square we divide the sum of squares by one less than the number of observations in the sample.

This concept of degrees of freedom is of immense importance in statistical methods. We have just seen that, having calculated the mean of the sample, we are left with one less than the total number of observations as the total number of degrees of freedom. But there is no reason why these should not be subdivided or *partitioned*. The principle of this may be illustrated by a homely example. Let us suppose that we have a number of articles to divide among ten people in such a way that we are free to give the first nine any number we wish, up to the total available. The remainder, if there is any, must then of course go to the tenth. Clearly, then, while we have discretion with regard to the number allotted to the first nine we have no discretion with regard to the tenth, and our possibilities of choice, that is our degrees of freedom, are only nine in all. But let us go further, and suppose that the people among whom we distribute the articles are themselves divided into two categories, say Englishmen and Scotsmen, and we have to divide the articles among the English and the Scots, as categories. Clearly, then, we have one degree of freedom among the categories, in that having given some to one, the other must have the remainder.

Suppose further that there are six English and four Scots. We will then have five degrees of freedom among the English, and three among the Scots, since in each case the last man among the six or four must have the remainder. We thus have five, plus three, plus one degree of freedom in all, giving us the total of nine for the sample. However they are partitioned, the sum of the groups of degrees of freedom must always add up to those of the sample.

In a large sample, the possibilities of thus partitioning the degrees of freedom become very numerous. For the moment we need not pursue the matter further, because its implications will be considered below,

in some of the examples. The principle opens up, however, the possibility of analyses of great flexibility and power.

To illustrate the arithmetical method of calculating the sum of squares and mean square we may first use the fuel consumption figures from Table 2 (page 25), carrying them to Table 5. These are small samples,

TABLE 5. CALCULATION OF SUM OF SQUARES AND MEAN SQUARE: FUEL CONSUMPTION OF LORRIES

1	2	3	4
3-ton petrol		5-ton diesel	
Column 2 of Table 2 with 7 deducted	Freshly calculated for this table	Column 3 of Table 2, with 14 deducted	Freshly calculated for this table
$x = C - 7$	x^2	$x = C - 14$	x^2
0·98 1·65 2·08 3·62 3·52 3·61 3·14	0·96 2·72 4·33 13·10 12·39 13·03 9·86	0·94 1·83 2·71 4·25 1·42 2·29 1·94	0·88 3·35 7·34 18·06 2·02 5·24 3·76
$S(\ )$ 18·60	56·39	15·38	40·65
$\bar{x}$	$\frac{18·60}{7} = 2·66$		$\frac{15·38}{7} = 2·20$
$\bar{C}$	$2·66+7 = 9·66$		$2·20+14 = 16·20$
$S(x^2)$	56·39		40·65
$\frac{S^2(x)}{n}$	$\frac{18·60^2}{7} = 49·42$		$\frac{15·38^2}{7} = 33·79$
$S(x-\bar{x})^2$	6·97		6·86
s^2	$\frac{6·97}{6} = 1·162$		$\frac{6·86}{6} = 1·14$
s	$\sqrt{1·16} = 1·08$		$\sqrt{1·14} = 1·07$

and so the work can be done directly, without grouping; but we can save some work in the arithmetic by deducting a convenient figure from each value of the fuel consumption, say 7 for the 3-ton lorries and 14 for the 5-tonners. This was not necessary in Table 2, because we were

merely adding the figures, but as they now have to be squared, it is worth doing.

In Table 5, the means will be seen to be the same as in Table 2 (page 25) less the figures deducted. The working should be clear from the table. The two root mean squares are seen to be practically the same, at just over 1 mile/gal. The deductions of 7 and 14 do not affect the estimation of the sum of squares. There is no further information which we can get from these figures.

Next, the sum of squares, mean square, and root mean square will be derived for the grouped sample used in Table 3 (page 26). The sum of squares can be calculated by two methods. The first is the standard one, and the second is a process of summation. The latter will be found to be less work, especially when a calculating machine is not available. Table 6, which shows the working, gives the two methods, the standard method (A), and the summation method (B). Anyone making the calculation is entirely free to use which method he prefers, at his choice.

Referring first to Table 6A, which sets out the standard method, columns 1 to 4 are the same as those columns in Table 3 (page 26), of which Table 6A is merely an extension. Column 5 of Table 6A is the fx column multiplied again by x—or the f column multiplied by x^2. We can now see why it is such a saving in time to use a working unit. If this was not done the figures in this column would be sixteen times as large as they are here. Then addition of column 5 gives $S(x^2)$ to find the mean square.

For the alternative method, refer to Table 6B. Here columns 1 to 3 are again columns 1, 2, and 3 of Table 3 (page 26), i.e. they again set out respectively the speed x and the frequencies. Column C1 is a repetition of column 4*a* of Table 3 (page 26), except that the frequency in the line for $x = 0$, i.e. the line for the working mean, is left out. It is the cumulative summation starting from the top of the table, of the frequencies in column 3. As before, its sum gives $S(fx)$. It has been headed C1 to indicate that it is the first cumulative column. Then in column 5, the C2 or second cumulative column, column C1, is itself summed cumulatively from the top in the same way. That is, taking the figures from column 5, the first line is 5, the next is $5+16=21$, the next is $21+42=63$, and so on. The line for the working zero is again left out. The bottom figure entered in this column should be the same as the sum of column C1, i.e. it is $S(fx)$.

Column C2 is then added up, doubled, and the last term of the column subtracted from the result. This then gives $S(fx^2)$. Thus

$$2S(\text{C2}) - S(\text{C1}) = S(fx)^2 \quad . \quad . \quad . \quad . \quad . \quad (8)$$

TABLE 6. CALCULATION OF SUM OF SQUARES AND MEAN SQUARE: VEHICLE SPEEDS

A. Standard Method

1	2	3	4	5
Speeds V, miles/h	x	f	fx	fx^2
56/60	11	5	55	605
52/56	10	11	110	1100
48/52	9	26	234	2106
44/48	8	52	416	3328
40/44	7	125	875	6125
36/40	6	213	1278	7668
32/36	5	353	1765	8825
28/32	4	431	1724	6896
24/28	3	299	897	2691
20/24	2	100	200	400
16/20	1	35	35	35
12/16	0	10	0	0
$S(\ )$		1660	7589	39779

$\bar{x} = \frac{7589}{1660} = 4{\cdot}5716$ in working units.

$= 4{\cdot}5716 \times 4 + 14$ in miles/h $= 32{\cdot}29$ miles/h

$S(fx^2) = 39779$

$\frac{S^2(fx)}{n} = \frac{34694{\cdot}5}{n}$

$S(x-\bar{x})^2 = 5084{\cdot}5$

$s^2 = \frac{5084{\cdot}5}{1659} = 3{\cdot}065$ in working units

$s = \sqrt{3{\cdot}065} = 1{\cdot}751$ in working units

$= 1{\cdot}7451 \times 4 = 7{\cdot}004$ miles/h

$s^2 = 3{\cdot}065 \times 16 = 49{\cdot}04$ in (miles/h)^2

Sheppard's adjustment $= \frac{16}{12} = \frac{1{\cdot}33}{47{\cdot}71}$ adjusted s^2 in (miles/h)^2

Adjusted $s = 6{\cdot}91$ miles/h

B. Summation Method

1	2	3	4	5
Speed V, miles/h	x	f	C1	C2
56/60	11	5	5	5
52/56	10	11	16	21
48/52	9	26	42	63
44/48	8	52	94	157
40/44	7	125	219	376
36/40	6	213	432	808
32/36	5	353	785	1593
28/32	4	431	1216	2809
24/28	3	299	1515	4324
20/24	2	100	1615	5939
16/20	1	35	1650	7589
12/16	0	10	—	—
$S(\ )$		1660	7589	23684

$2S(\text{C2}) = 2 \times 23684 = 47368$

$S(\text{C1}) = 7589$

$S(fx^2) = 39779$

Rest of working as for Part A

The working for this is shown underneath Table 6B. This method will be seen to be a saving in work. The derivation of this process is given in Section 16.3 in Chapter 16.

Then, the mean in working units $\bar{x}$ is worked out below the main table. Below that again the sum of squares is worked out by substitution in eqn. 6, in whichever of the two forms is preferred. The sum of squares when worked out from eqn. 6 is the difference of two relatively large quantities, and so the second term in the equation—i.e. either

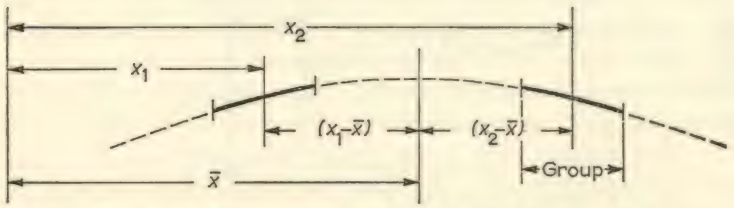


Fig. 7. Diagram illustrating the bias caused by grouping

$\bar{x}S(x)$ or $S^2(fx)/n$—must be taken to a fair degree of accuracy, preferably to one place further than $S(fx^2)$, which has, of course, been derived direct from the figures, and so is exact. Then below this again s^2 and s are worked out. All this calculation is done in working units, and that is usually all that is needed, because the normal use of the tables is to find the 'statistics' which will be described later. The root mean square s, is however, finally given in miles per hour, by multiplying by 4, the working unit.

The process of grouping introduces a bias into the result for the mean square, because the distribution in a group not actually covering the mean itself is a sloping line. The principle of this is shown in Fig. 7. In this figure the short lengths of full line show two groups with the same numerical value of $(x-\bar{x})$ on either side of the mean, and together with the dashed line they show part of the curve of a normal distribution. In the ordinary way, there will tend to be a greater number of values of $(x-\bar{x})$ on the inner side, nearer to the mean, than on the side further away from the mean. In working out the mean these differences will largely cancel out, but in working out the mean square they will not, because the differences are squared, so with grouped data the mean square will tend to be too large.

This may be compensated for by making *Sheppard's adjustment*, which is done by subtracting one-twelfth of the square of the unit of grouping. In this case, then, we would make Sheppard's adjustment by

deducting $16/12 = 1{\cdot}333$ from the calculated value of the mean square, as shown in Table 6. The adjustment should not, however, be made when the tests of significance, to be described later, are to be done.

Example 5.1. Using the data of Example 4.1, and continuing that example, calculate the mean square and root mean square of the accident death rate for the two states Connecticut and Rhode Island for the years 1947 to 1959.

Example 5.2. Using the data of Example 2.1, and continuing Example 4.2, calculate the mean square and root mean square of the annual winter rainfall for the Thames basin for the period 1883 to 1959.

*Example 5.3.** In Gumbel's method of flood prediction, the range of expected maximum floods, F_m, during a period of T years, is given by the equation

$$F_m = F_1 + aT' \pm \frac{aT''}{\sqrt{n}}$$

In this expression, F_1 and a are quantities based on the mean and standard deviation of floods recorded during a period of n years, and T' and T'' are functions of T whose values have been tabulated.

When T is taken as 5 years, this equation becomes

$$F_m = \bar{F} + 0{\cdot}72s \pm \frac{1{\cdot}75s}{\sqrt{n}}$$

In this expression, $\bar{F}$ is the mean flood recorded during the period of n years, and s is the root mean square of the floods recorded during that period. The last term is a correction for the error introduced by the fact that $\bar{F}$ and s are estimated from the sample of n years.

The table below gives data for a series of floods recorded between 1938 and 1951. From these data, estimate F_m for a period of 5 years.

Year	Flood, cusec	Year	Flood, cusec	Year	Flood, cusec
1938	6550	1943	6110	1948	5563
1939	4180	1944	4600	1949	4964
1940	3950	1945	4550	1950	5940
1941	3000	1946	2350	1951	4850
1942	3280	1947	6471		

* Adapted, with modified notation, from 'Hydro-Electric Engineering', Vol. I, edited by J. Guthrie Brown (Blackie, 1957).

CHAPTER 6

Significance

In statistical methods, no attempt is made to provide a 'proof' of anything, in the popular sense of a word which is loosely used by the many amateurs who deal in figures, most of whom have axes of some sort to grind. This loose use of the word 'proved' by amateurs is responsible for the fact that there is an element of truth in the saying attributed to Disraeli or Mark Twain, 'There are three kinds of lies: lies, damned lies, and statistics'! Engineers—particularly highway engineers—suffer very frequently from this kind of amateur, and must be especially careful. While figures cannot lie if they are accurate in themselves, they can still be so used as to be very untruthful.

It often happens that there are wide fluctuations between figures for the same operation in succeeding years, and the amateur, basing his opinion partly on what he wants to find, and partly on the notorious *post hoc, ergo propter hoc* argument, says that the difference 'proves his point'. As an example of this one may cite the claim often made that because the introduction of the 30 miles/h speed limit was followed by a fall in the number of deaths on the roads the following year, the case for the limit was 'proved'. But in random figures of this sort wide fluctuations are always possible. If the trend of the figures for road deaths is examined over a number of years before the limit was introduced we find that there were wide fluctuations during these years. Moreover, at the same time as the limit was introduced a number of other things happened which might have had some effect on the death rate, such as pedestrian crossings, an intense propaganda campaign, various improvements in the vehicles and the roads, and so forth. It can hardly be said with any real confidence that any particular one of these measures stopped accidents unless more evidence can be found. They may all have had some effect, or one may have caused the whole fall; we do not know. Thus the most that we can say is that the case for

the limit cannot be accepted with complete confidence on the figures for Great Britain taken alone.

In one of the books on road accidents—not one written by a traffic engineer—the figures of Table 7 were given for the number of deaths

TABLE 7. ROAD DEATHS IN A CITY

Year	No. killed
1946	24
1947	16
1948	22
1949	25
1950	8
1951	13

in a certain city during the years 1946 to 1951. Late in 1948 a great increase took place in the number of convictions of motorists, as a result of a campaign for the strict enforcement of the law. It was claimed in the book that these figures 'proved' the value of this campaign.

Now admittedly a cursory examination of these figures certainly does make it appear as if the prosecutions must have had a profound effect. But we are dealing with exceedingly complex events, which are very rare, statistically speaking. A road death, it must be remembered, occurs once in about ten million vehicle-miles. That there are about seven thousand a year is because there are about nine or ten million vehicles on the road, which travel between them an astronomical number of miles a year, and also that there are about fifty million or so people using the roads on foot, and several millions on cycles. It is also partly a matter of chance whether an accident results in death, or no injury at all. Thus considering these figures objectively, we cannot feel sure merely from inspection, that they indicate a real effect, or that they are due to random chance fluctuations. We need some means of testing this.

This kind of operation can be done by means of what are called *tests of significance.* These are a means of detecting whether a difference between two sets of observed data is due to a real difference between the populations from which they are drawn, and not to random chance differences between samples drawn from the same population. It is obviously highly improbable that two samples drawn from the same population would be identical, and the tests of significance are used, in effect, to determine whether these differences between samples can be accounted for by chance.

If the difference between the sets of data being investigated is said to

be *significant*, it means that the two samples are likely to have been drawn from different populations. The term 'significant' is an abbreviation for 'significantly different from zero'. The usual practice is to conclude that the difference is significant if the probability of getting one as large or larger than that observed is less than one in twenty, or 0·05, or 5%, all of which are different ways of expressing the same thing. This figure is called the *level of significance*, or *significance level*. Both terms are in very common use, and it is often said that a difference is *significant at the* 5% *level*, or whatever is the chosen level. The use of 5% is not a hard-and-fast rule, but one which has been found as a result of many years study and experience to be a good working compromise. It involves the risk that a true hypothesis will be rejected once in twenty trials.

There is, however, no reason why another level of significance should not be used, and it may be varied according to the importance of the consequences of the rejection of the hypothesis. Some discussion of this is made below. The level of significance should always be stated in any published work, so that another worker reading the report can judge for himself. That is, it should never be stated that 'the difference was significant' but that 'the difference was significant at the j% level'.

The more common tests of significance are based on the principle that if a variate is normally distributed with variance σ^2, then the means of a number of random samples, each of n observations, are themselves normally distributed with a variance σ^2/n. This again is a principle of fundamental importance. It theoretically depends on the variate being normally distributed, but in practice it applies to many distributions which are not normal, because even when individual observations are not normally distributed, the means of moderate sized samples of them are much more nearly normally distributed. This greatly extends the availability of the tests. The use of the principle may be extended further by transforming variates which cannot be normally distributed, in the way already described (page 21).

Even if truly random samples are taken from a really normal distribution, small ones will not usually look as if they are normal when they are plotted. The larger the samples are, the more nearly normal will they appear. Thus a sample of ten observations may show little sign of any regularity of distribution, while one of 100 may show clear signs of it. But the tests of significance allow for these effects. The differences between samples of different sizes is illustrated by Tippett.[5]

CHAPTER 7

The Principles of the Tests of Significance

The tests of significance are all related to each other, and so a general description will be given first. This should greatly help in an understanding of the tests themselves. Their application will be described in detail in later chapters. The relationship between the tests is explained by Mather.[4]

1. *The Normal Deviate*

This depends on a knowledge of the variance of the population, which is not usually available. Thus it is not of general use, but as it is the basis of the other tests, it will be described first. Denoting the normal deviate by c, we have that

$$c = \frac{|d|}{\sigma} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

In this expression, d denotes the deviation of any one observation from the mean, as shown in Fig. 5, which, it will be remembered, also shows σ. Dividing d by σ eliminates scale, and so makes the normal deviate c quite general. The probability of getting any particular value of the normal deviate has been calculated, and tables are published in many books. An abridged table is given at the end of this book, in Table E. While c itself is not often used, its tabulation is helpful.

As an approximation, it is useful to remember that the probability of getting a deviation equal to the standard deviation—i.e. a normal deviate of 1—is nearly one in three, of twice the standard deviation is nearly one in twenty, of two-and-a-half times the standard deviation is one in one hundred, and of three times the standard deviation is about one in four hundred. So a normal deviate of 2 very nearly represents the 0·05 or 5% level of significance; and one of 2·5 very nearly

represents the 0·01 or 1% level, which is sometimes used when tests of greater delicacy are needed.

2. *The t Test*

We do not usually know the variance of the population, and all the information we may have about it is the estimate derived from the sample, i.e. the mean square, the square root of which is denoted by s. So a function t is used. This is the same as the normal deviate in principle, except that the deviation d is divided by the root mean square, instead of the standard deviation, so that

$$t = \frac{d}{s} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (10)$$

This function is used for estimating the significance of single deviations from the mean. The denominator has as many degrees of freedom N as the mean square from which the root mean square s has been derived, and t is tabulated in terms of this N. When N is greater than about 30, t and the normal deviate are almost equivalent, and a table of c can be used for t for large samples without serious error. Thus for more than 30 degrees of freedom in the denominator, a t of 2—more exactly 1·96—will be significant at the 0·05 level, and one of 2·5 will be significant at the 0·01 level. An abridged table of t is given in Table F at the end of the book.

3. *The Variance Ratio Test*

This is the most flexible of the tests of significance, because it can be applied when there are any number of degrees of freedom in both numerator and denominator. If we wanted to make a comparison between several means we would get a term with several degrees of freedom in the numerator, whilst t has only one. The test is done by finding the ratio between two mean squares, and the probability of getting various values of this ratio has been tabulated. Fisher, who was the first to apply the test, calculated and tabulated a function z, which is half the natural logarithm of the ratio between the mean squares. Snedecor tabulated the direct ratio, which is usually called the *variance ratio*, and he denoted it by the letter F, as a compliment to Fisher. The use of the function in this form is becoming general.

As there are two sets of degrees of freedom, it is impossible to use a single table for the variance ratio, and it is usual to give tables for various suitable levels of significance. At the end of the book, Tables G1 to G3 are given for levels of significance of 0·20, 0·05 and 0·01. These levels should serve most practical purposes.

4. *The χ^2 Test of Goodness of Fit*

The χ^2 test compares observations with those expected from some known or suspected hypothesis. It is thus of very frequent practical use. To perform the test, we compute

$$\chi^2 = S\left[\frac{(O-E)^2}{E}\right] \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

In this expression, E is the frequency expected from the hypothesis, and O is the frequency observed. The quantity given by eqn. 11 has approximately the so-called χ^2 distribution. In effect χ^2 is analogous to a ratio between two mean squares, in which the denominator is obtained from the hypothesis, and so has an infinite number of degrees of freedom. The numerator may have any number of degrees of freedom, depending on how it has been derived.

There are many forms of this expression, some of which will be given when the test is described in detail. The function is sometimes described as 'chi-square'.

CHAPTER 8

The t Test

We have seen in the last chapter that the t test is based on dividing a deviation by the estimate of the standard deviation—the root mean square s—derived from the sample. The general expression for t is

$$t = \frac{|d|}{s} \quad . \quad . \quad . \quad . \quad . \quad (10) \text{ repeated}$$

In many cases we use t to test the significance of a difference between the means of two samples, using for our hypothesis the assumption that both samples are derived from the same population, of which the variance will be σ^2, and the mean square estimated from the combined samples will be s^2. We use the principle, already stated (page 40) that the means of samples of n observations taken from a normal distribution are themselves normally distributed with variance σ^2/n, and our estimate of this is s^2/n.

From this, if we state the deviation as the difference between a mean of a sample and that of the population, the general form of eqn. 10 becomes

$$t = \frac{|\bar{x}-\mu|}{s/\sqrt{n}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

In this expression, $\bar{x}$ is the mean of a sample, and μ that of the population.

If we test the difference between the means of two samples numbered 1 and 2, and give their properties numbers according to that of the sample, so that the number of observations in the first sample is n_1 and that in the second n_2, then t is derived from the expression

$$t = \frac{(\bar{x}_1-\bar{x}_2)}{s}\sqrt{\left(\frac{n_1 n_2}{n_1+n_2}\right)} \quad . \quad . \quad . \quad . \quad . \quad (13)$$

and

$$s^2 = \frac{S(x_1-\bar{x}_1)^2+S(x_2-\bar{x}_2)^2}{n_1+n_2-2} \quad . \quad . \quad . \quad . \quad . \quad (13a)$$

The number of degrees of freedom N, with which to look up in the table the probability of getting such a value of t, is given by

$$N = n_1+n_2-2 \quad . \quad . \quad . \quad . \quad . \quad . \quad (14)$$

One mean has been calculated from each sample, so that one degree of freedom has been lost in each, making two lost in all. If it is desired to test whether a single mean differs significantly from zero, then $\bar{x}_2$ in eqn. 13 becomes zero, n_2 is neglected, and eqn. 14 becomes $N=n-1$. If both samples are the same size n, the term in the bracket in eqn. 13 becomes $n/2$.

The root mean square s used in eqn. 13 is derived by adding together the mean squares for the two samples, each divided by its own degrees of freedom, because we are assuming, as the hypothesis to be tested, that both samples come from the same population. This depends on the principle, already mentioned, that the means of samples derived from a normal distribution are themselves normally distributed with the variance σ^2/n. Thus in estimating the variance of the difference, we again divide the variance of the samples, combined, by the combined degrees of freedom, and again take its square root.

If each had the same estimated variance s^2 as they would if the hypothesis were true, the variance of the mean of the first sample on the principle already quoted above would be s^2/n_1 and that of the second sample would be s^2/n_2. In combining them in eqn. 13, to allow for our now having the combined sample of (n_1+n_2) observations, we divide the difference between the means by

$$\sqrt{\left(\frac{s^2}{n_1}+\frac{s^2}{n_2}\right)} = s\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = s\sqrt{\left(\frac{n_1+n_2}{n_1 n_2}\right)}$$

which gives us the divisor in eqn. 13.

Taking first a very simple example, we may test the figures of Table 7 (page 39) for significance. It will be remembered that this table gave the road deaths in a certain city for two periods each of three years before and after a campaign for more prosecutions of motorists and of greater severity in the courts began late in 1948. The working is set out in Table 8. In this the sums of squares are calculated directly, because the samples are so small. The rest of the working should be clear from the table, and from the foregoing description.

The value of t obtained, looked up in the table of t for four degrees of freedom, has a probability of more than 0·1. In the table, and in the

following tables, the probability is denoted by p. From a more detailed table given by Fisher[6] it appears to be about 0·4, and so the difference

TABLE 8. THE t TEST: ROAD DEATHS IN A CITY

Before campaign of severity	After campaign of severity
24 16 22	25 8 13
$S(\ )$ 62	46
$\bar{x}_1 = \frac{62}{3} = 20{\cdot}67$ $S(x_1-\bar{x}_1)^2 = 34{\cdot}67$	$\bar{x}_2 = \frac{46}{3} = 15{\cdot}33$ $S(x_2-\bar{x}_2)^2 = 152{\cdot}66$

The combined sum of squares $= 34{\cdot}67 + 152{\cdot}66 = 187{\cdot}33$

So $$s^2 = \frac{187{\cdot}33}{4} = 46{\cdot}83$$

and $$s = \sqrt{46{\cdot}83} = 6{\cdot}84$$

$$t = \frac{(\bar{x}_1-\bar{x}_2)}{s}\sqrt{\frac{3}{2}} = \frac{(20{\cdot}67-15{\cdot}33)}{6{\cdot}84} \times 1{\cdot}225 = 0{\cdot}96$$

For 4 degrees of freedom $p > 0{\cdot}1$.
Hence the difference is not significant.

would be found twice in about five trials from genuinely random figures. We cannot, therefore, conclude with any confidence that the prosecutions reduced the deaths. The samples are, of course, very small, so that the random variations mentioned above could be expected to play a large part.

We can learn some lessons from this example. The author quoted was making a case for more punishments for motorists, and, as the figures were convenient for this, she accepted them as they stand. But she should have begun by enquiring into their accuracy. At that time, about 8% of road deaths in the country were in accidents in which no motor vehicle was present, so some of this type might have been included in the figures. Next, she should have used the number of fatal accidents, not the number of deaths. It is a matter of chance whether one, or many, deaths occur in a single accident.

Next, she could have enquired whether the campaign of prosecutions was the *only* change in road conditions in the city during the period covered, which was actually one in which all sorts of measures might have been taken which would be expected to reduce accidents. Thus

street lighting might have been improved after the wartime blackout. This is known to reduce accidents. Yet it seems that no study of such things was made, so the figures as they stand are quite valueless and I have only used them because they are arithmetically convenient, and as illustrating the traps into which we can fall.

We may also consider what level of significance should be used in such a case. The consequences of prosecutions and severe penalties on individuals subjected to them may be so serious that it is reasonable to ask what level of proof that the public good is served by prosecutions we should demand, before inflicting such severe consequences on individuals. We have also to consider the possible social effects if severe penalties are inflicted on inadequate grounds. The consideration of this is outside the scope of this book, and all that is being done is to point out the presence of the problem. Putting the matter in other words, we may ask whether we are justified in inflicting severe penalties on individuals on an off-chance of saving accidents. If not, then we might conclude that we could reasonably ask for a higher level of significance, and greater confidence that our figures disclose a real effect, than that afforded by the conventional 0·05 level, but we can hardly consider the matter further here.

Turning next to a more complicated example, with grouped data and a much larger sample, we may use the speed counts in Dorset, given in Table 1 (page 9). Some of these were taken in places subject to the 30 miles/h speed limit, and others in places not subject to that limit, but where one had been demanded by some such body as a Parish Council or Road Safety Committee. The stretches of road covered by these counts are roughly similar in character. The demand for the limit is usually made in places which have been recently developed, or which have always been borderline cases. The counts within the limits were taken in similar places. One of them was on a very wide road, with a suburban-type development, but, it might be added, a low accident rate. It is, in fact, the type of place where motorists would not see the need for a speed limit, and so it allows a good test of the effect a speed limit has on the speed of traffic when not affected by other circumstances.

In this example, the two sets of timings, those within the speed limits, and those in the other places, are added together and are tested to find if there is any significant difference between them. The working is given in Table 9. In this table, the fx and fx^2—or C1 and C2—columns are not included. Those for the second sample have already been given in Table 3 (page 26) and so should not need to be repeated. The working follows the same lines as the last example. Here the number of degrees of freedom for t is so large as to be virtually infinite, and so the

probability can be taken from the table of the normal deviate. It is about 0·09, which gives some indication that the difference between the means is real, but it would not normally be said to be significant. The term *indication* is used in the sense of meaning that we may suspect a real

TABLE 9. THE t TEST: TIMINGS OF VEHICLE SPEEDS

1	2	3
Speed V, miles/h	(*a*) Within speed limits	(*b*) Places where speed limit demanded
56/60	–	5
52/56	1	11
48/52	6	26
44/48	21	52
40/44	37	125
36/40	116	213
32/36	184	353
28/32	219	431
24/28	141	299
20/24	56	100
16/20	11	35
12/16	1	10
$S(f)$	793	1660
$S(fx)$	3529	7589
$S(x^2)$	17527	39779
$\bar{x}$	4·4502	4·5716
$\bar{V}$, miles/h	31·8	32·3
$S(x^2)$	17527	39779
$S^2(x)/n$	15704·7	34694·5
$S(x-\bar{x})^2$	1822·3	5084·5

Combined sum of squares $= 1822{\cdot}3 + 5084{\cdot}5 = 6906{\cdot}8$

$$s^2 = \frac{6906{\cdot}8}{2451} = 2{\cdot}818$$

Thus $s = 1{\cdot}679$ and $\bar{x}_1 - \bar{x}_2 = 0{\cdot}1214$

$$t = \frac{0{\cdot}1214}{1{\cdot}679}\sqrt{\left(\frac{793 \times 1660}{2453}\right)} = 1{\cdot}67$$

N is virtually infinite. p is about 0·09 (from Table E).
Hence there is a difference indicated but it is not significant (see also Table 9*a*).

difference, though it is not significant, and that we cannot have full confidence it is a real difference.

It may sometimes happen that the two mean squares are significantly different. This is more fully explained in the next chapter, and for the

moment it should be enough to mention that this is so for large samples if the result of dividing the larger mean square by the smaller is a quantity appreciably larger than unity. If the t test has shown that the difference between the means is decidedly either significant or not significant, as the case may be, this should not matter very much. But if the difference is nearly, or only just, significant, a more sensitive form of the test should be done. It will be seen that the case of Table 9 fits these conditions. The ratio of the mean squares is 1·33, and the test as carried out has indicated that the difference is nearly significant, so that the modified test is advisable. In the example of Table 8 this form of the test need not be made, because the mean squares are not significantly different.

In this form of the test, instead of pooling the sums of squares and then dividing by the combined degrees of freedom to obtain s, as in eqn. 13*a*, we must calculate each estimated sample variance separately by dividing the sum of squares for each of the samples by its own degrees of freedom; thus $s_1^2 = S(x_1 - \bar{x}_1)^2/(n-1)$ for sample 1, and similarly for sample 2. Then we again divide s_1^2 by n_1 and s_2^2 by n_2 and add these two results. The difference between the means is then divided by the square root of this quantity to find t. Thus

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \quad \ldots\ldots (14a)$$

But when we do this we cannot use eqn. 14 to find N, the number of degrees of freedom with which to look up the new value of t. We must find N from the following expression

$$N = \frac{(\theta+v)^2(n_a-1)(n_b-1)}{\theta^2(n_b-1)+v^2(n_a-1)} \quad \ldots\ldots (14b)$$

In this expression, $\theta = s_a^2/s_b^2$ and $v = n_a/n_b$. In these, s_b is the *bigger* root mean square, and n_b its number of observations. The samples have been given letters here to distinguish them, because either sample 1 or sample 2 may have the larger mean square. For arithmetical convenience the smaller mean square is divided by the larger in this case. When both samples are the same size n, this expression becomes

$$N = \frac{(\theta+1)^2(n-1)}{(\theta^2+1)} \quad \ldots\ldots (14c)$$

It is seldom necessary to calculate N from eqn. 14*b* because it always gives a number bigger than (n_b-1), though smaller than (n_a+n_b-1).

In the example of Table 9 it is not necessary because we can look upon N as infinite, but in the example of Table 9*a*, which gives the modified

TABLE 9A. MODIFIED ANALYSIS OF TABLE 9

The table proper is as shown in the upper half of Table 9, the following replacing the analysis in the lower half of Table 9.

$$s_a^2 = s_1^2 = \frac{1822{\cdot}3}{792} = 2{\cdot}301 \qquad s_b^2 = s_2^2 = \frac{5084{\cdot}5}{1659} = 3{\cdot}065$$

$$\frac{s_a^2}{n_a} + \frac{s_a^2}{n_b} = \frac{2{\cdot}301}{793} + \frac{3{\cdot}065}{1660} = 0{\cdot}002905 + 0{\cdot}001846 = 0{\cdot}00474$$

$$|\bar{x}_b - \bar{x}_b| = 0{\cdot}1214$$

$$t = \frac{0{\cdot}1214}{\sqrt{0{\cdot}00475}} = \frac{0{\cdot}1214}{0{\cdot}0689} = 1{\cdot}76$$

$$s_a^2 = 2{\cdot}301 \qquad s_b^2 = 3{\cdot}065 \qquad \theta = 0{\cdot}7508$$
$$n_a = 793 \qquad n_b = 1660 \qquad v = 0{\cdot}4777$$
$$(\theta + v) = 1{\cdot}2285$$

$$(\theta + v)^2 = 1{\cdot}509 \qquad (n_a - 1) = 792 \qquad (n_b - 1) = 1659$$
$$\theta^2(n_b - 1) = 935{\cdot}3 \qquad v^2(n_a - 1) = 180{\cdot}7$$

$$N = \frac{1{\cdot}509 \times 792 \times 1659}{935{\cdot}2 + 180{\cdot}7} = 1777$$

$0{\cdot}1 > p > 0{\cdot}05$; nearly significant (from Table E)

A more detailed table included by Fisher[6] gives p as almost 0·08. Lindley and Miller[7] give p as 0·078.

Note: To find the probability of the normal deviate in Table 1 given by Lindley and Miller[7], look up the value of x the same as that of c and read off $\Phi(x)$, then $p = 200[1 - \Phi(x)]$.

method applied to the case of Table 9, the work has been done to illustrate the method. Table 9*a* does not include the working out of the means and sums of squares, because this is the same as in Table 9. We see that the difference is now slightly more nearly significant than was found from Table 9.

This raises another matter of importance in our studies. Even if the difference is significant, we are still entitled to ask if it is important. The two questions are not the same. If we say the difference is significant we mean that we would not expect it to occur by chance in random samples—no more than that. But a difference may be statistically significant, yet unimportant in practice. In the example of Table 9, the difference between the means amounts to only about ½ mile/h, so it does not amount to very much. It does not appear, then, from these two sets of figures, that the behaviour of the motorists timed was very materially affected by the speed limits, especially because the mean speed within the limited lengths was nearly 2 miles/h over the limit.

So far as we have gone, however, we have not enough data to enable us to get any further with our conclusions, though there is, in fact, more information available for further tests, which will be described in the next chapter.

Example 8.1. Using the data of Example 4.1, and continuing the analysis of that example and of Example 5.1, test the significance of the difference between the means of the annual death rates per hundred million vehicle-miles in the states of Connecticut and Rhode Island.

Example 8.2. The table below shows the frequencies of lives of road surface dressings laid in 1948 and 1949, using tar and gravel. Test whether the mean lives of the two sets of dressings are significantly different.

Year dressed	Life in years													
	1	2	3	4	5	6	7	8	9	10	11	12	13	> 13
1948	2	3	9	13	9	9	4	8	3	–	1	1	1	1
1949	1	4	6	15	21	20	7	5	1	4	2	–	–	–

CHAPTER 9

The Analysis of Variance

Hitherto, the data have not been divided into more than two sets, but it may often happen that there are more than two to be compared. In the example we have just considered, there were two sets of timings, but these sets were themselves subdivided, because they were made up of a number of timings in different places, taken at different times. In some of these the conditions were such that speeds would not be expected to be high, while in others they might be high, because there was no obvious reason for a speed limit. It may thus be of some interest to compare the various timings to see if they agree reasonably well with each other. Such comparisons as these can be done by means of the *analysis of variance*.

This test enables any number of samples to be compared, in contrast with the *t* test, which only enables two to be compared. It is the most general form of these types of analysis.

In the analysis, we arrange the samples to be compared in columns, as we have done in the preceding work, and it is usual to call each column an *array*. Then all the arrays are added together to form the *grand array*, which is thus the whole volume of the data collected together. Its mean is called the *grand mean*. The grand array is usually put on the right of the table.

Before discussing the arithmetical method, it is advisable to discuss the principles of the analysis. This may be represented diagrammatically, as in Fig. 8. In this diagram, each array is represented by a lens-shaped hatched area, which must be imagined to contain all the observations in the array. The larger area on the right stands for the grand array. The mean of an array is denoted by $\bar{x}_a$, and if it were needed to express the mean of a particular array it could be given a suffix, thus $\bar{x}_{a1}$, though this is rarely necessary. The grand mean—the mean of all the

observations taken together in the grand array—is denoted by $\bar{X}$. All means are shown on the diagram by the symbol ⊕.

First, one single observation—chosen at random—will be discussed. It is here taken in the fourth array, and is shown by a heavy dot. This observation may be considered as representative of all the observations.

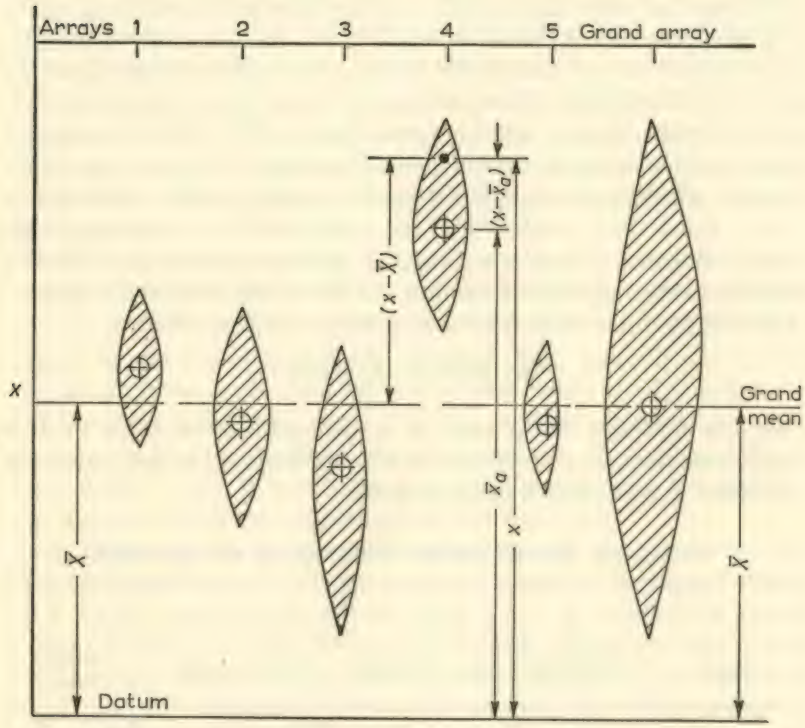


Fig. 8. Diagrammatic representation of the analysis of variance table

It is distant $(x-\bar{X})$ from the grand mean, and $(x-\bar{x}_a)$ from the mean of its own array. These distances are shown on the diagram, and dimensioned. The mean of each array will be distant $(\bar{x}_a-\bar{X})$ from the grand mean.

Each of these distances has its own sum of squares. The first, the largest of these, containing all the others, is the sum of squares of all the observations, taken together in the grand array, i.e. it is the sum of squares of the grand array, $S(x-\bar{X})^2$. It measures the scatter of all the observations about the grand mean. It has $(n-1)$ degrees of freedom, n being as usual the total number of observations. This general

scatter may be divided into two separate scatters. First, each individual array has its own sum of squares, which measures the scatter of its own set of observations about its own array mean. This term is, for any array, $S_a(x-\bar{x}_a)^2$. As mentioned above, it may sometimes be necessary to distinguish the arrays by suffixes, but this is not necessary here. In this term $S_a()$ denotes the summation for any one array. Then the sum of all these terms, $SS_a(x-\bar{x}_a)^2$, measures the scatter of all the observations, each in its own array, and each about its own array mean. In other words, it measures the scatter of the observations within the arrays. If there are m arrays, this quantity has $(n-m)$ degrees of freedom. Finally, we have another term $S[n_a(\bar{x}_a-\bar{X})^2]$. This measures the scatter of the array means about the grand mean. In it n_a denotes the number of observations in any given array and $\bar{x}_a$ is the mean for that array. The term is usually called the scatter between the arrays. It has $(m-1)$ degrees of freedom. Then the degrees of freedom of the two terms—the scatter within the arrays and the scatter between the arrays—together equal the total degrees of freedom; or in symbols,

$$(n-m)+(m-1)=(n-1)$$

We may represent this process by a small table as in Table 10. It is most important that this should be fully understood. Its derivation will be found in Section 16.4 in Chapter 16.

TABLE 10. PRINCIPLES OF THE ANALYSIS OF VARIANCE

1	2	3	4	5
Scatter	Degrees of freedom	Sum of squares	Mean square	Variance ratio
Within arrays	$(n-m)$	$SS_a(x-\bar{x}_a)^2$	$\frac{SS_a(x-\bar{x}_a)^2}{(n-m)}=A$	$\frac{A}{B}$ if $A>B$ or $\frac{B}{A}$ if $B>A$
Between arrays	$(m-1)$	$S[n_a(\bar{x}_a-\bar{X})^2]$	$\frac{S[n_a(\bar{x}_a-\bar{X})^2]}{(m-1)}=B$	
Total	$(n-1)$	$S(x-\bar{X})^2$		

Table 10 shows that, to discover whether the various samples represented by the arrays can reasonably be thought of as coming from the same population, represented by the grand array, we find the ratio of the mean square of the scatter between the arrays to that within the arrays. To do this, each of the sums of squares is divided by its own degrees of freedom, to find its mean square. Then the larger of the

mean squares is usually divided by the smaller, because the quantity thus found, which is called the *variance ratio*, is tabulated as greater than unity, and this is looked up in the table of this ratio (Snedecor's F). In the tables, N_1 is the number of degrees of freedom of the larger mean square, and N_2 that of the smaller. Fisher used a function z which is half the difference between the natural logarithms of the mean squares. The modern tendency is to use F. The process may be expressed in words by saying that we are finding out whether the scatter of the array means about the grand mean is in proportion to the scatter within the arrays themselves which would be expected from random sampling. However, if the 'within arrays' mean square is significantly the larger one, the result is suspect. There has either been a mistake in the arithmetic, or the data are biased, or their distribution is seriously non-normal.

This will be illustrated by examples, the first of which is given in Table 11. This table shows the individual timings within the speed limits. We notice that the grand array is the same as set (a) in Table 1 (page 9) and Table 9 (page 48). It lists, in fact, the individual timings from which the set in these tables had been made up. The upper part of the main table is worked out by the methods described above, as far as the line for $\bar{V}$. This latter line does not play any part in the calculation, which is wholly done in working units, but it is put in to give some idea of the mean speeds in miles per hour.

It should be noted—as a very useful check on the arithmetic—that $SS_a(fx)$ should equal $S(fx)$ for the grand array, i.e. the sum of the line for $S_a(fx)$, namely 3529, should equal $S(fx)$ as worked out for the grand array; also 3529 is the sum worked out in the same line of column 8 of the main table. This is also true for $S(fx^2)$, but not for the remaining four lower lines of the table, a fact emphasized by separating columns 7 and 8 with a double vertical line alongside these lines. The sum of squares for the grand array is worked out at the foot of column 8.

Then on the left below the main table the sums of squares of the arrays are set out in a column, and added, to find $SS_a(x-\bar{x}_a)^2$. On the right of this, this term is deducted from $S(x-\bar{X})^2$, the sum of squares of the grand array, which gives $S[n_a(\bar{x}_a-\bar{X})^2]$. As an alternative, and as a check on the calculation, this quantity can be worked out directly from the means and sums of squares of the arrays, which is done in a separate table at the bottom of Table 11. We see that the difference is very small, the figure from the main table being 318·3, and that from the check calculation 318·2. These tables, and some of the following ones, are thus to some extent self-checking; as the arithmetic is fairly complex, that is very useful, and full advantage should be taken of it.

To help in the understanding of the tables, in this and all subsequent corresponding examples, the figures carried to the analysis of variance table are trebly underlined.

The analysis of variance table corresponds to Table 10 (page 54) and is below the main table. The sums of squares worked out in the main table, and trebly underlined, are carried to it. Each is then divided by its own number of degrees of freedom to find the mean square; thus $318{\cdot}3/5 = 63{\cdot}64$. Then the larger mean square is divided by the smaller: thus $63{\cdot}64/1{\cdot}91 = 33{\cdot}4$, which is the required variance ratio. This is in the right hand of the column. It will be remembered that the larger mean square is divided by the smaller one. Usually, though not invariably, the larger mean square has the fewer degrees of freedom.

The variance ratio is then looked up in the tables to find the probability of getting the calculated value by random sampling. It is not possible to give the full range of probabilities for all number of degrees of freedom. The variance ratio is, as it were, three-dimensional. It is usual to give tables for a small number of probability levels. In many books, as in this one, tables are given for levels of 0·20, 0·05 and 0·01 (Tables G1, G2, G3, pages 136–138). It is not often necessary to find the level of probability very exactly. In the tables, the number of degrees of freedom of the numerator, those of the larger mean square N_1, are in the top line; and the number of degrees of freedom N_2 of the denominator, those of the smaller mean square, in the left-hand column. This is always the rule, no matter whether the number of degrees of freedom of the numerator is larger or smaller than that of the denominator. The variance ratio is the reading at the intersection of the column for N_1 and the row for N_2.

In this example N_1 is 5; N_2 is 787, and so is virtually infinite. Thus in Table G3 (page 138), if we look down the column for $N_1 = 5$ to the bottom line for $N_2 = \infty$, we find a variance ratio of 3·0. Ours is 33·4, so that the probability is very much less than 0·01. If we refer to a more detailed table, given by Lindley and Miller,[7] we find that for the 0·001 probability level the ratio is 4·1; so in our example the probability is very small indeed, and the differences between the arrays are very highly significant. This would not be unexpected from the large differences between the mean speeds of the various counts. It seems probable that the driver pays little attention to the imposed limit, but adjusts his speed to the conditions of the actual length of road he is traversing.

A similar result was found from the counts in the places where the limit had been demanded, and the pattern in those places was very similar to that within the limits. It might also be added that the appearance of the roads in the two sets of timings was in general very similar.

TABLE 11. ANALYSIS OF VARIANCE:
TIMINGS OF VEHICLE SPEEDS

1	2	3	4	5	6	7	8
Array	1	2	3	4	5	6	Grand array
Speed	f	f	f	f	f	f	f
52/56	1	—	—	—	—	—	1
48/52	5	1	—	—	—	—	6
44/48	4	5	1	—	—	—	21
40/44	7	9	1	1	4	—	37
36/40	27	24	15	16	4	1	116
32/36	29	54	23	38	12	1	184
28/32	34	47	46	50	27	21	219
24/28	9	37	32	47	24	27	141
20/24	3	12	10	21	15	20	56
16/20	1	3	—	2	9	6	11
12/16	—	—	—	—	1	1	1
$S_a(f)$	120	192	128	181	96	76	793
$S_a(fx)$	613	852	520	901	427	216	$SS_a(fx) = S(fx) = 3529$
$\bar{x}_a$	5·1083	4·4375	4·0625	4·9779	4·4479	2·8421	$\bar{\bar{x}} = 4{\cdot}4502$
$S_a(fx^2)$	3439	4192	2292	4799	2115	690	$SS_a(fx^2) = S(fx^2) = 17527$
$\bar{x}\,S_a(fx)$	3131·4	3780·8	2112·5	4485·1	1899·3	613·9	$S^2(fx)/n = 15704{\cdot}7$
$S(x-\bar{x}_a)^2$	307·6	411·2	179·5	313·9	215·7	76·1	$S(x-\bar{\bar{x}})^2 = 1822{\cdot}3$
s^2	2·585	2·153	1·413	1·744	2·271	1·015	
$\bar{V}$, miles/h	34·4	31·8	30·3	33·9	31·8	25·4	

$s^2 = \frac{1822{\cdot}3}{792} = 2{\cdot}301$

$s = 1{\cdot}52$

Array	$S_a(x-\bar{x}_a)^2$
1	307·6
2	411·2
3	179·5
4	313·9
5	215·7
6	76·1
$SS_a(x-\bar{x}_a)^2$	1504·0

$S(x-\bar{\bar{x}})^2 = 1822{\cdot}3$

$SS_a(x-\bar{x}_a)^2 = 1504{\cdot}0$

$S[n_a(\bar{x}_a-\bar{\bar{x}})^2] = 318{\cdot}3$

Analysis of variance

Scatter	Sum of squares		N		Mean square	Variance ratio	p
Between arrays	$S[n_a(\bar{x}_a-\bar{\bar{x}})^2]$	318·3	$(m-1)$	5	63·64	33·4	<0·01 Highly significant
Within arrays	$SS_a(x-\bar{x}_a)^2$	1504·0	$(n-m)$	787	1·91		
Total	$S(x-\bar{\bar{x}})^2$	1822·3	$(n-1)$	792			

Check calculation

Array	1	2	3	4	5	6	
$\lvert\bar{x}_a-\bar{\bar{x}}\rvert$	0·6581	0·0127	0·3877	0·5277	0·0023	1·6081	
n_a	120	192	128	181	96	76	
$n_a(\bar{x}_a-\bar{\bar{x}})^2$	51·97	0·03	19·24	50·40	0·00	196·53	$S[n_a(\bar{x}_a-\bar{\bar{x}})^2] = 318{\cdot}2$

Since the t test showed that there was a slight, though not significant, difference between the two sets of timings, we might think that the limit had some effect, but only a slight one, although the driver possibly does in the main tend to adjust his speed to the conditions prevailing, irrespective of the presence or absence of a speed limit. However, we also see from Fig. 3 that there is some indication that the effect of the limit is to reduce the higher speeds, and this effect may account for the difference between the two sets of timings.

It must, however, be pointed out that while this gives us some indication of the actual speeds, it does not give us any indication whatever of the relative danger or safety in the various places. We have found that the limits do not very materially reduce the speed, but we have not found out whether this makes for danger or not. Speed and danger are not necessarily related, statistically speaking, and their relationship would have to be the subject of a further investigation, and until that is done, the question is quite open. We must not take anything for granted, however convinced we may feel, *a priori*, that reduced speed makes for safety.

In the foregoing analysis, we found that there are very wide differences between the arrays, but we have no means of carrying the matter further, and finding out how the differences arise. We can suspect that there are also wide differences between the stretches of road on which the timings were taken, and this certainly appeared on the ground to be true, but there is no objective way of measuring this. So apart from noting that the differences exist we cannot find out anything more about them, and in this case we have extracted all the available information about the speed counts.

The method is, however, capable of further development with suitable data. Quite frequently we can further subdivide the data into two or more families of arrays which have some feature in common, and we may suspect that there are differences between the families; or we may want to find out whether the differences are between the families and the arrays, or where they are. It is possible to extend the analysis of variance to test this. We may denote the properties of the family by the suffix f, so that $\bar{x}_f$ will denote the mean of a family and $S_f()$ a summation for a family. The family will really be a sub-grand array, with its own sub-grand mean, and the combination of the families will be the grand array.

When this is so, our sum of squares between arrays, that is $S[n_a(\bar{x}_a-\overline{X})^2]$, will now be further divisible in its turn, into two more terms. The first of these will measure the scatter within the arrays of a family, the second the scatter between the means of the families. This

is expressed in Table 12, which is similar to Table 10 (page 54), but extends and subdivides the middle line of that table. In this it will be taken that there are m arrays and j families. The derivation of this follows the same lines as that for Table 10 given in Section 16.4 in Chapter 16.

TABLE 12. ANALYSIS OF VARIANCE: SUBDIVISION OF THE SCATTER BETWEEN ARRAYS INTO SCATTER BETWEEN AND WITHIN FAMILIES

Scatter	Degrees of freedom	Sums of squares
Between arrays of the same family	$(m-j)$	$SS_f[n_a(\bar{x}_a-\bar{x}_f)^2]$
Between families	$(j-1)$	$S[n_a(\bar{x}_f-\overline{X})^2]$
Total	$(m-1)$	$S[n_a(\bar{x}_a-\overline{X})^2]$

The method is now clear. We first carry out the analysis as before from Table 10, to test whether the differences between arrays are greater than we would expect from random sampling. We then go on to test the differences between the families, using the method outlined in Table 12. The next example will show the method.

The example (Table 13) is taken from a series of tests of cement/sand briquettes. In these tests, nine experimenters each made and tested three briquettes. Six of them used sand A, and the three others sand B. We now have two possibilities; first, differences between the experimenters, and second differences between sands. In Table 13 the first line shows the experimenters, their numbers serving as the numbers of the columns. The second line shows the sand they used. The next three lines the tensile strengths they obtained, each less 300 lb/in^2 to reduce the size of the figures. Experimenter 6 had one reading of 295 lb/in^2, which is entered as -5; and experimenter 8 had one reading of exactly 300, which is entered as 0. The next line gives the results added up for the arrays, and these in turn are summed at the right of the table in column 10 to give the grand array. The next four lines follow the usual practice to find the means, and the sums of squares, for the arrays, and these are again summed at the right of the table, in column 10.

The bottom four lines of the main table, from the line for $|x_a-\overline{X}|$ down to the bottom, inclusive, are again a check on the arithmetic. The third and fourth lines from the bottom, giving the various values of $n_a(\bar{x}_a-\overline{X})^2$, are of course, similar to those of the check calculation in Table 11. The bottom two lines are also similar, but give the calculation

TABLE 13. ANALYSIS OF VARIANCE: TENSILE TESTS OF CEMENT/SAND BRIQUETTES

Experimenter	1	2	3	4	5	6	7	8	9	10 (Grand array)
Sand used	A	A	A	A	A	A	B	B	B	
x = Tensile strength −300 lb/in²	196	150	145	165	275	154	150	71	111	
	122	135	117	156	175	151	103	16	89	
	55	40	40	36	116	−5	64	0	75	
$S_a(x)$	373	325	302	357	566	300	317	87	275	$2902 = SS_a(x)$
$\bar{x}_a$	124·33	108·33	100·67	119·00	188·67	100·00	105·67	29·00	91·67	$107·48 = \bar{X}$
$S_a(x^2)$	56325	42325	36314	52857	119706	46542	37205	5297	25867	$422438 = SS_a(x^2)$
$\bar{x}_a S_a(x)$	46376	35208	30401	42483	106785	30000	33496	2523	25208	
$S_a(x-\bar{x}_a)^2$	9949	7117	5913	10374	12921	16542	3709	2774	659	$69958 = SS_a(x-\bar{x}_a)^2$
$\lvert\bar{x}_a-\bar{X}\rvert$	16·85	0·85	6·81	11·52	81·19	7·48	1·81	78·48	15·81	
$n_a(\bar{x}_a-\bar{X})^2$	852	2	139	398	19775	168	10	18476	750	$40570 = S[n_a(\bar{x}_a-\bar{X})^2$
$\lvert\bar{x}_a-\bar{x}_f\rvert$*	0·83	15·17	22·83	4·50	65·17	23·50	30·23	46·44	16·23	
$n_a(\bar{x}_a-\bar{x}_f)^2$	2	690	1564	61	12740	1657	2741	6470	790	

* For $\bar{x}_f$ see table on facing page.

Total for grand array	For sand A				For sand B			
	$n_a(\bar{x}-\bar{x}_f)^2$	$S(x-\bar{x}_a)^2$	$S_fS_a(x) = 2223$	$S_fS_a(x^2) = 354069$	$n_a(\bar{x}_a-\bar{x}_f)^2$	$S(x-\bar{x}_a)^2$	$S_fS(x) = 679$	$S_fS_a(x^2) = 68369$
	2	9949	$*\bar{x}_f = 123·50$	$\bar{x}_fS_fS_a(x) = 274541$	2741	3709	$*\bar{x}_f = 75·44$	$\bar{x}S_fS_a(x) = 51227$
$SS_a(x^2) = 422438$	690	7117		$S_f(x-\bar{x}_f)^2 = 79528$	6470	2774		$S_f(x-\bar{x}_f)^2 = 17142$
$\bar{X}SS_a(x) = 311911$	1564	5913		$S_fS_a(x-\bar{x}_a)^2 = 62816$	790	659		$S_fS_a(x-\bar{x}_a)^2 = 7142$
$S(x-\bar{X})^2 = 110527$	61	10374		$S_f[n_a(\bar{x}_a-\bar{x}_f)^2] = 16712$				
	12740	12921			10001	$7142 = S_fS_a(x-\bar{x}_a)^2$		$S_f[n_a(\bar{x}_a-\bar{x}_f)^2] = 10000$
	1657	16542						
	16714	$62816 = S_fS_a(x-\bar{x}_a)^2$						

$SS_f[n_a(\bar{x}_a-\bar{x}_f)^2] = 16712 + 10000 = 26712$ to Table B below

Analysis of variance

Table	Item	Sum of squares		N	Mean square	Ratio	p	
A	Within arrays	$SS_a(x-\bar{x}_a)^2$	69958	18	3887	1·304	> 0·2	Not significant
	Between arrays	$S[n_a(\bar{x}_a-\bar{X})^2]$	40569	8	5071			
	Total	$S(x-\bar{X})^2$	110527	26				
B	Within arrays	$SS_a(x-\bar{x}_a)^2$	69958	18	3887	$3·565 = t^2$	$0·1 > p > 0·05$	Nearly significant
	Between arrays of same sand	$SS_f[n_a(\bar{x}_a-\bar{x}_f)^2]$	26712	7	[3816]	$t = 1·89$		
	Between sands	$S[n_a(\bar{x}_f-\bar{X})^2]$	13857	1	13857			
	Total	$S(x-\bar{X})^2$	110527	26				

of the sum of the squares of the deviations of the array means from the family means. Each sand is taken as a family. The two calculations of the figures agree well.

The figures from column 10, in the first three lines, are used to find the sum of squares for the grand array, in the first, or left-hand, panel of the set of tables below the main table. This is headed 'Total for grand array'. Then the relevant trebly underlined figures are carried to the upper part of the analysis of variance table, marked A in the first column. From this it will be seen that the probability of getting the deviations found is more than 0·2, or one in five, and so is not significant. We may then go further, and do the more sensitive test to find out the effect of differences between the sands, taking each sand as a family.

The next part of the calculation is in the two panels in the middle table, headed 'For sand A' and 'For sand B', respectively. They repeat the usual process of finding the sum of squares, in this case for the two families for the two sands. Both should be clear from what has gone before. Sand A will be one family and sand B another.

Then these figures are carried in their turn to the lower Analysis of Variance table, marked B in the first column. This is similar to the top one A of the two tables, with the difference that the scatter between the arrays has been subdivided according to Table 12 (page 59). The figures in the second and third lines in Table B for the two sums of squares, 26712 and 13857, add up to 40569, which is the sum of squares in column 10 of the main table, 40570, except for a difference of 1, owing to minor arithmetical errors. We then calculate the mean squares within arrays, and between sands by dividing as usual by the degrees of freedom, and then divide the larger by the smaller, i.e. we divide 13857 by 3887. As there are two sands, i.e. two families, there is only one degree of freedom for their mean square, and so their variance ratio is t^2. Taking its square root, then, gives us t for 18 degrees of freedom. This works out at 1·89, giving a probability of between 0·1 and 0·05, which is very nearly significant. It is not essential to do this, as the variance ratio may be looked up directly, but the t table gives us a little more detail, and so is more convenient. The variance ratio for one degree of freedom in the numerator, that is when $N_1 = 1$, is t^2.

Thus the second stage of the analysis shows that although the first analysis did not find that the differences were significant, the particular cause of variation between the sands was more important than appeared at first. It will also be noticed from the lower analysis that the mean square for the scatter within arrays of the same sand and between arrays of the same sand, 3887 and 3816, are almost equal, giving a variance

ratio of almost unity. This means that the variations between experimenters using the same sand are of the same order as would be expected from the scatter of the individual experimenter's results. Thus the whole, or practically the whole, of the differences between the arrays is associated with the differences between the sands. By this more sensitive test we have been able to find out where the difference lies.

Example 9.1. In a method study of fencing work, the following timings were made of tying chain link fencing to line wires (figures in hundredths of minutes). Make an analysis of variance, and find whether the times taken for the various rows are significantly different.

	Times								
Top row	30	30	45	42	40	39	41	34	
Middle row	29	52	45	44	40	42			
Bottom row	41	37	50	48	53	56	30	30	23

For these figures, my thanks are due to Mr. J. H. H. Wilkes, County Surveyor of Somerset, and to Mr. L. P. Vincett, formerly Method Study Engineer in Somerset.

Example 9.2. The results shown in the following table were obtained in a comparative test between two laboratories, in which three observers in each laboratory each tested three specimens of the same road tar for viscosity. Make an analysis of variance, and test whether the results obtained in the two laboratories are significantly different.

Laboratory	1			2		
Observer	A	B	C	X	Y	Z
Viscosities (seconds at 30°C)	216	208	195	198	210	212
	220	212	203	198	209	204
	214	203	187	195	208	206

Note: If the results obtained by the two laboratories are found not to be significantly different, test whether observers C and X taken together are significantly different from the rest taken together.

CHAPTER 10

Regression

In the analysis of variance considered in Chapter 9 the arrays could be put into any order without affecting the analysis, but it is not unusual for us to have a set of data in which each observation is of the type (x, y), so that there are two variates. We may then want to find out whether they are connected by some relationship. It is possible to do this by an extension of the analysis of variance, which is called *regression*. The most striking difference in the arrangement of the analysis for the purpose of regression is that the position of the arrays is decided by their values of a variate x; otherwise the basic principle is the same.

The simplest type of expression by which the variates can be connected is the equation of a straight line:

$$y = a + bx \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (15)$$

If we are going to carry out a regression analysis to find out whether the variates are connected by such an expression, the first step must be to decide to which of our variates to give the symbol y. Often this is easy, because it is not unusual for one of them to be under the control of the experimenter, or to be decided for him in some way, and so to be less liable to chance variations, other than those of ordinary experimental error. A frequent example of this type of variate is time. The object of the analysis may be to find the variation of some function or other with time. In such cases the experimenter may take the intervals of time to suit his convenience. On the other hand, he may be examining the variations of some quality over the years—one of our examples will be to analyse the variations of fatal accidents over the years—and then the time scale will be decided for him. Another example might be a study of the strength of concrete cubes with a varying water/cement

ratio, in which case the latter quality would be under his control, and he would choose it to suit the range of his experiments. In such cases this type of variate is called the *independent variate*, and it is given the symbol x. The other is then called the *dependent variate*, and given the symbol y. The expression found in the analysis—if one is found—is called the *regression equation.*

The reason for this distinction in name is that a regression equation is not quite the same thing as an ordinary algebraic equation. With a regression equation of the form $y = f(x)$, we are not entitled to use ordinary algebraic methods to transform it into one of the form $x = f(y)$. Very often this distinction is of little importance, because the form $x = f(y)$ may be of little interest to us. Thus we would hardly find an equation for years in terms of fatal accidents of much practical meaning! A regression equation does not express an invariable value of y for any given value of x, as an algebraic equation does. It expresses the average value of y for that value of x, and the analysis by which it is derived depends on a process of minimizing the squares of the deviations from the *regression line*—which is the line resulting from the plotting of the solution of the regression equation—and it is only valid in the form derived from that minimization. If it is wanted in the reverse form, the variates must be changed over, and the minimizing done afresh. To transform the equation by algebraic means would thus give a false result.

It is, of course, possible for both the variates to be subject to chance variations, so that there is no true independent variate. In such cases the arrangement of the variates must be chosen to suit the form in which the expression is needed. This is explained again below, in connection with one of the examples.

For regression purposes we express the equation for the straight line, eqn. 15, in the form

$$Y = a + b(x - \bar{x}) \quad . \quad . \quad . \quad . \quad . \quad . \quad (16)$$

In this, Y means the average value of y derived from the regression equation. This practice is followed because in many cases there will be many observed values of y to correspond with any one value of x, which in its turn may only be a mean value. The distinction between Y and y is important in the working which follows.

The process of derivation of the regression equation is illustrated in Fig. 9. This is similar in principle to Fig. 8, which was used to explain the derivation of the analysis of variance, except that most of it is given in the form of a dot diagram, for greater clearness. In Fig. 9, diagrams (*a*) and (*b*) show a small number of observations. In (*a*) they are just

considered as a grand array, and the dashed lines pass through the means $\bar{x}$ and $\bar{y}$, which are those of all the observations taken together. Then, with reference to the y axis, an observation taken at random will be distant $(y-\bar{y})$ from the mean, as shown in Fig. 9*a*. To this corresponds a sum of squares $S(y-\bar{y})^2$, which corresponds to the sum of squares of the grand array in the analysis of variance. This sum of squares has $(n-1)$ degrees of freedom.

If there really is an association between the variates represented by a straight line, we can reduce the size of the sum of squares, $S(y-\bar{y})^2$, by taking it with regard to a sloping straight line representing the equation we want. Turning now to Fig. 9*b*, in which the sloping line has been inserted, and expressing as Y the value of y found from the line that corresponds to some given value of x, we have the original distance $(y-\bar{y})$ reduced to $(y-Y)$. In other words, if we have an observation with coordinates x_j and y_j^*, its distance from the grand mean, $(y_j-\bar{y})$, will now be divided into two; thus

$$(y_j-Y_j)+(Y_j-\bar{y})=(y_j-\bar{y})$$

Here Y_j is the corresponding value to x_j obtained from the regression line. Each of these types of distance will have its own sum of squares, $S(y-Y)^2$ and $S(Y-\bar{y})^2$. Our problem, then, is to adjust the values of the coefficients a and b in eqn. 16 so as to minimize the sum of squares $S(y-Y)^2$.

When this is done (see Section 16.5 in Chapter 16), the two coefficients are derived from the expressions

$$a=\bar{y} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (17)$$

$$b=\frac{S[y(x-\bar{x})]}{S(x-\bar{x})^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad (18)$$

The numerator in eqn. 18 is a very important quantity called a *deviation cross-product*. When it is divided by $(n-1)$ the resulting quantity is called the *covariance*. It has three identical forms (see Section 16.6 in Chapter 16):

$$S[(x-\bar{x})(y-\bar{y})] \quad \text{or} \quad S[x(y-\bar{y})] \quad \text{or} \quad S[y(x-\bar{x})]$$

It is calculated from the expression (see Section 16.6 in Chapter 16)

$$S[y(x-\bar{x})]=S(xy)-n\bar{x}\bar{y}$$

$$=S(xy)-\frac{S(x)S(y)}{n} \quad . \quad . \quad . \quad . \quad (19)$$

The covariance may be negative or positive.

* Here j is not necessarily an integer.

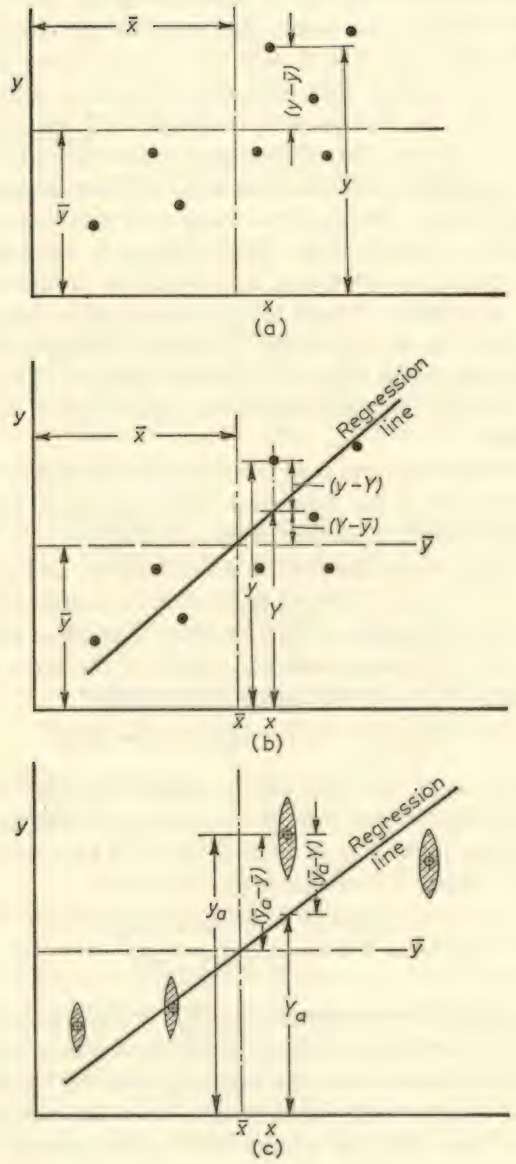


Fig. 9. Diagram to illustrate the principles of regression

When there is a much larger number of observations, so that grouping becomes necessary, there will be an extra sum of squares, though the principle remains the same. Referring to Fig. 9*c*, this again represents the data as in Fig. 8, with the arrays shown as lens-shaped areas. The observations may, however, be related either to definite values of x, after the fashion of the example in Table 13 (page 60), or to the mean of a group. Thus they might be observations for such things as concrete specimens with a definite series of water/cement ratios. This type of observation—though for notched steel specimens—is used in a later example in Table 16 (page 76). On the other hand they may have scattered x values, in which case they would be grouped with definite values of x, as in Table 17 (page 80). In either case the principle remains the same, and Fig. 9*c* may apply to either. The symbol $\oplus$ used, as before, may thus apply either to a chosen value of x, or to the mean of a group, though it is always called the array mean in the explanation which follows.

The total sum of squares is $S(y-\bar{y})^2$, which represents the scatter of all the observations, in the y direction, about the grand mean. It can be divided into two parts. The first of these is $SS_a(y-\bar{y}_a)^2$, in which $S_a()$ denotes as before a summation for a single array, and $\bar{y}_a$ is the mean of an array. This sum of squares represents the scatter of the observations each about the mean of its own array. The other sum of squares is $S[n_a(\bar{y}_a-\bar{y})^2]$. This represents the scatter of the array means about the grand mean. It is derived from the expression*

$$S[n_a(\bar{y}_a-\bar{y})^2] = S[\bar{y}_a S(y)] - \bar{y}S(y) \quad . \; . \; . \quad (20)$$

This second sum of squares will be reduced by the introduction of the regression line, so that it in its turn can be divided into two parts. The first of these is the sum of squares removed by the regression line, i.e. $S(Y-\bar{y})^2$. This is derived from the expression

$$\begin{aligned} S(Y-\bar{y})^2 &= bS[y(x-\bar{x})] \\ &= b^2S(x-\bar{x})^2 \quad . \; . \; . \; . \; . \quad (21) \end{aligned}$$

The second part is a remainder $S[n_a(\bar{y}_a-Y_a)^2]$, which represents the deviation of the array means about the regression line. In this term Y_a is the value of Y derived from the regression line for the same value of x as that for the particular array mean. Then the ratio of this latter term to the term for the scatter within the arrays, $SS_a(y-\bar{y}_a)^2$, tests the deviations from regression. The coefficients a and b are derived from eqns. 17 and 18.

* Further derivations are not included in Chapter 16, because enough are included to enable a reader to work out others for himself if he wishes to do so.

Having derived b, we must test its significance, and this can be done in two ways. First by means of a t test. For this, the estimate of the variance, which is s^2, is needed, and

$$s^2 = \frac{1}{n-2}S(y-Y)^2 \quad . \; . \; . \; . \; . \; . \quad (22)$$

and in this expression

$$S(y-Y)^2 = S(y-\bar{y})^2 - bS[y(x-\bar{x})] \quad . \; . \; . \quad (23)$$

Then

$$t = \frac{b\sqrt{S(x-\bar{x})^2}}{s} \quad . \; . \; . \; . \; . \; . \; . \; . \quad (24)$$

This will have $(n-2)$ degrees of freedom.

This process can also be expressed in the form of an analysis of variance table, as in Table 14, and then the regression mean square divided by the error mean square gives t^2 for testing the significance of b. The two forms of t are the same.

TABLE 14. REGRESSION: ANALYSIS OF VARIANCE TABLE FOR TEST OF SIGNIFICANCE OF b

Item	Sum of squares	Degrees of freedom	Mean square	t^2
Regression	$bS[y(x-\bar{x})]$ [or $b^2S(x-\bar{x})^2$]	1	$bS[y(x-\bar{x})]$ [or $b^2S(x-\bar{x})^2$]	$\frac{bS[y(x-\bar{x})]}{s^2}$
Remainder or error	$S(y-Y)^2$	$n-2$	$\frac{S(y-Y)^2}{n-2} = s^2$	
Total	$S(y-\bar{y})^2$	$n-1$		

A full understanding of the process of regression is very important, and so three examples will be given, in increasing order of arithmetical complexity. The first is a double one with ungrouped data, with time as the independent variate. There is only one observation for each value of x, which simplifies the arithmetic. Table 15 shows two tables of the accidental deaths in England and Wales over a number of years. The figures are taken from the appropriate reports issued by the Registrar-General. Table 15A gives the deaths in motor-vehicle traffic accidents, and Table 15B other accidental deaths. It should be noted that the road deaths are not the same as those given in many tables, as these almost always include deaths in incidents not involving motor vehicles. These latter deaths are here included in Table 15B. An analysis will be

done to find if there has been any trend in the numbers of either or both of these types of death over the years, and a further analysis will be done to see if the two trends, if there are any, are different.

Table 15 gives the number of motor vehicle traffic deaths over the years 1946 to 1959 in Column 2 of Part A, though 3500 has been deducted from the total to reduce the size of the figures, as has been done before. The other accidental deaths are in Column 5 of Part B, with 9000 deducted, for the same reason. Columns 3 and 6 are summations from the top of the column to find $S(xy)$, in a manner similar to the summation method described previously (page 27). Thus we have actually taken x as zero in 1960, and numbered it backwards over the years, and so have inverted the table. It is not essential to do this, but in a table of this type it makes it easier to add figures for later years, if we desire to follow up the trend when later figures become available. The column for x has not been included in the table, as it is unnecessary. Columns 4 and 7 are respectively y_1^2 and y_2^2 worked out directly by squaring y_1 and y_2.

In this case, x is a variate, and so we need $\bar{x}$ and $S(x-\bar{x})^2$. The numbers form the series 1, 2, 3, 4, . . . n, and for such a series we have available the general expressions

$$S(x) = \tfrac{1}{2}n(n+1) \quad \text{and} \quad S(x^2) = \tfrac{1}{6}n(n+1)(2n+1)$$

Using these expressions, it follows from eqn. 4 that

$$\bar{x} = \frac{n+1}{2} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (25)$$

and from eqn. 6 that

$$S(x-\bar{x})^2 = \tfrac{1}{12}n(n^2-1) \quad . \quad . \quad . \quad . \quad . \quad (26)$$

These functions, which are common to both tables, are worked out on the right, alongside the main table. Under this again are worked out $\bar{y}_1$ and $S(y_1-\bar{y}_1)^2$ from Table 15A, and $\bar{y}_2$ and $S(y_2-\bar{y}_2)^2$ from Table 15B, by the usual methods. They result in large figures, so that a calculating machine has to be used to get the necessary accuracy.

Under the main table, the deviation cross-products are worked out, by substitution in eqn. 19, and from them, the regression coefficients b_1 and b_2 are calculated. These turn out to be negative because the tables have been inverted in time. In the left-hand panel under the main table, marked A, the regression equation has been worked out directly, with the negative coefficient, and then it has been reversed into the form we need it for use, underneath, with the final form doubly underlined. In the right-hand panel the reversal has been done directly. From the

TABLE 15. REGRESSION: ACCIDENTAL DEATHS IN ENGLAND AND WALES 1946–1959*

1	2	3	4	5	6	7
Year x	*A.* Motor vehicle traffic accidents			*B.* Accidents other than motor vehicle traffic accidents		
	Deaths -3500 y_1	Summation for $S(xy_1)$	y_1^2	Deaths -9000 y_2	Summation for $S(xy_2)$	y_2^2
1946	307	307	94249	955	955	912025
1947	338	645	114244	1893	2848	3583449
1948	6	651	36	305	3153	93025
1949	264	915	69696	465	3618	216225
1950	634	1549	401956	679	4297	461041
1951	892	2441	795664	1407	5704	1979649
1952	471	2912	221841	991	6695	982081
1953	746	3658	556516	1435	8130	2059225
1954	947	4605	896809	2919	11049	8520561
1955	1308	5913	1710864	2224	13273	4946176
1956	1439	7352	2070721	2268	15541	5143824
1957	1327	8679	1760929	1954	17495	3818116
1958	1866	10545	3481956	2364	19859	5588496
1959	2452	12997	6012304	2265	22124	5130225
$S(\ )$	12997	63169	18187785	22124	134741	43434118

$n=14 \qquad \bar{x} = \frac{n+1}{2} = \frac{15}{2} = 7{\cdot}5$

$S(x-\bar{x})^2 = \frac{n}{12}(n^2-1) = 227{\cdot}5$

A: $\bar{y}_1 = \frac{12997}{14} = 928{\cdot}3571$; $S(y_1-\bar{y}_1)^2 = 18187785 - 12065858 = 6121927$

B: $\bar{y}_2 = \frac{22124}{14} = 1580{\cdot}286$; $S(y_2-\bar{y}_2)^2 = 43434118 - 34962241 = 8471877$

A. For motor vehicle accidents

$S[y_1(x-\bar{x})] = 63169 - 12997 \times 7{\cdot}5 = -34308{\cdot}5$

$b_1 = \frac{-34308{\cdot}5}{227{\cdot}5} = -150{\cdot}807$

$b_1\, S[y_1(x-\bar{x})] = 5173948$

$Y_1 = \bar{y}_1 + b_1(x-\bar{x}) = 928{\cdot}36 - 150{\cdot}8(x-7{\cdot}5)$
$= 928{\cdot}36 + 1131{\cdot}0 - 150{\cdot}8x$
$Y_1 = 2059{\cdot}4 - 150{\cdot}8x$

Reversing with $x = 0$ at 1945

$Y_1 = 928{\cdot}36 + 150{\cdot}8(x-7{\cdot}5)$
$= -202{\cdot}6 + 150{\cdot}8x$
$Y_1 = 3297 + 150{\cdot}8x$

B. For other accidents

$S[y_2(x-\bar{x})] = 134741 - 22124 \times 7{\cdot}5 = -31189{\cdot}0$

$b_2 = \frac{-31189{\cdot}0}{227{\cdot}5} = -137{\cdot}095$

$b_2\, S[y_2(x-\bar{x})] = 4275841$

Reversing $Y_2 = 1580{\cdot}3 + 137{\cdot}1(x-7{\cdot}5)$
$= 1580{\cdot}3 - 1023{\cdot}25 + 137{\cdot}1x$
$Y_2 = 9557{\cdot}0 + 137{\cdot}1x$

* The figures are taken from the Registrar-General's Reports.

(*Continued overleaf*)

TABLE 15 (*Continued*)

Analysis of variance

Coefficient	Item	N	Sum of squares	Mean square	Ratio
b_1	Regression function	1	$b_1 S[y_1(x-\bar{x})]$	5173948	$t^2 = 65{\cdot}5$
	Deviation from regression	12	$S(y_1-Y_1)^2$	947979	78998 $t = 8{\cdot}09$
	Total	13	$S(y_1-\bar{y}_1)^2$	6121927	

$p < 0{\cdot}01$: b_1 highly significant

Coefficient	Item	N	Sum of squares	Mean square	Ratio
b_2	Regression function	1	$b_2 S[y_2(x-\bar{x})]$	4275841	$t^2 = 12{\cdot}2$
	Deviation from regression	12	$S(y_2-Y_2)$	4196036	349670 $t = 3{\cdot}50$
	Total	13	$S(y_2-\bar{y}_2)^2$	8471877	

$p < 0{\cdot}01$: b_2 highly significant

Testing the difference between b_1 and b_2

$d = |b_1 - b_2| = 150{\cdot}8 - 137{\cdot}1 = 13{\cdot}7$

b_1 deviation mean square = 78998

b_2 deviation mean square = 349670

$$V_d = \frac{V_{y_1} + V_{y_2}}{S(x-\bar{x})^2} = \frac{428668}{227{\cdot}5}$$

$$s_d = \sqrt{\frac{428668}{227{\cdot}5}} = 43{\cdot}4$$

$$t = \frac{13{\cdot}7}{43{\cdot}4} = 0{\cdot}32 \;\&\; N = 12+12 = 24$$

$p = 0{\cdot}70$: not significant

two equations we see that 3297 motor-vehicle deaths, and 9557 other accidental deaths, would have been expected from the figures to have occurred in 1945, increasing at an annual rate of 150·8 for the motor vehicle deaths, and 137·1 for the other accidental deaths. It is again possible to do a check calculation by adding columns for Y worked out from the same equations, and then working out $(y-Y)$ and $(y-Y)^2$, summing the latter column. This has not been done here.

Having found the regression coefficients, we next test them for significance. This is done in the two analysis of variance tables in the continuation of the table on p. 72. These give us a value of t of 8·09 for b_1 and of 3·50 for b_2. In both cases these are very highly significant for 12 degrees of freedom, and we can thus conclude that we do not need to find a more elaborate expression than the straight line to represent the figures.

We can then go on to test whether the difference between the two regression coefficients is significant, i.e. whether the figures show that the deaths from the two types of accident have increased at different rates. This is done at the bottom left of the main table. The difference d between the two coefficients is $|b_1 - b_2| = |150{\cdot}8 - 137{\cdot}1| = 13{\cdot}7$. It is not necessary to take account of the sign. In both cases, the estimate of the variance of the coefficient b, which we assume to be the true regression coefficient for both sets of data, is the deviation mean square from the corresponding analysis of variance table. For b_1 this is 78998, and for b_2 it is 349670.

The assumption we are testing is that the two coefficients are the same, i.e. the difference is zero. If this assumption is justified, we are entitled to pool the two mean squares. Then, assuming that the true rate of increase of both types of accidents is b, we can estimate the 'true' b as $\frac{1}{2}(b_1+b_2)$. Also, if y_1 and y_2 are independent, the estimate of the variance—denoted by V_b—of b, the 'true' rate, is

$$V_b = \frac{V_{y_1} + V_{y_2}}{4S(x-\bar{x})^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (27)$$

V_{y_1} and V_{y_2} are the estimates of the variance of y_1 and y_2.

So, pooling the two mean squares of b_1 and b_2 on the assumption we are testing that the b's are the same, the estimate of the variance V_d of d is

$$V_d = V_{b_1} + V_{b_2} = \frac{V_{y_1} + V_{y_2}}{S(x-\bar{x})^2} \quad . \quad . \quad . \quad . \quad . \quad (28)$$

This differs from eqn. 27 in that it is the estimated variance of the difference, and not of the 'true' coefficient b. Then s_d, the root mean

square of the difference, is the square root of V_d, and t will be d/s_d. The working for this test is at the bottom of the table, on the left, and from this we see that the difference is definitely not significant, having

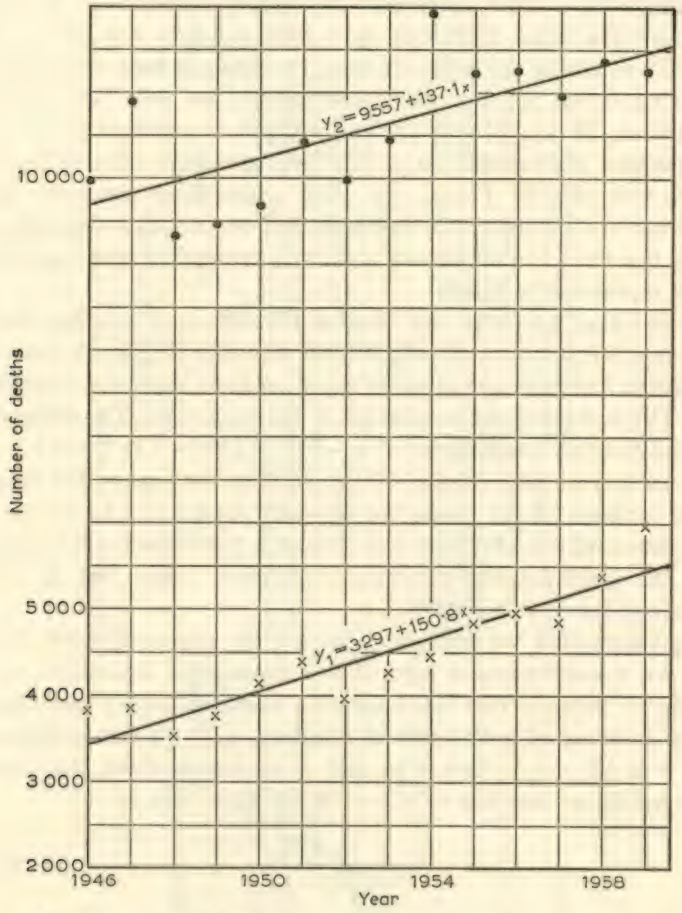


Fig. 10. Accidental deaths in England and Wales from 1946 to 1959, with regression lines

● Deaths in accidents other than motor vehicle traffic accidents.
× Deaths in motor vehicle traffic accidents.

a probability of about 0·70. This means that there is no reason to think that the two trends are different, surprising though that conclusion may be.

The observations, and the two regression lines, are plotted in Fig. 10. This shows the fairly close fit of the line for the motor vehicle

accidents, and the slightly less close fit of the line for the other accidents, as would be expected from the analysis of variance.

The next example is taken from a series of impact tests on notched steel specimens, and is given in Table 16. In this table, there are several sets of observations for each value of x, and while there are not enough to make grouping necessary, they do cause the work to be of greater arithmetical complexity. The effect to be investigated was whether the radius at the foot of the notch had any effect on the impact figure obtained. The radius R is clearly at the choice of the experimenter, and so it must be made the independent variate x; the impact value therefore must be made y. Six specimens were tested for each of four different radii, and the results, and the analysis, are given in Table 16. As grouping is not necessary, the individual readings are set out in the first six lines of the table, with 400 deducted from each figure to reduce the arithmetic. The next line shows the notch radius, and the lines below give the usual summations. The columns of notch radii, columns 2 to 5, are the arrays. Their sums, forming those of the grand array, are at the right of the table, in column 6.

On the extreme right of the table, relevant figures from the summations are set out, and b is calculated. Below that is the t test to test the significance of b. It will be seen to be very highly significant, with a probability of less than 0·01. Then the analysis of variance table tests whether the deviations from the regression line are significant. It is based on the method described on pages 68 & 69, using eqn. 20 and 21 on those pages. This was not necessary last time as each array was a single observation. The analysis shows that while the regression coefficient is highly significant, the deviations from regression are not significant. Here $N_1 = 20$ and $N_2 = 2$, because the 'within arrays' has the larger mean square. The figures for N_1 are not given very closely, but it is clear that p is much more than 0·2. The figures, and the regression line, are plotted in Fig. 11.

The check calculation is again done at the bottom right of the table, in which $S[n_a(\bar{y}_a - Y)^2]$ is worked out directly from the values calculated from the equation, and it does not differ materially from the value in the analysis of variance table, the two figures being 695 and 694·5.

The next example is one of much greater complexity. It illustrates the use of data in which both x and y are grouped, and also the use of transformed units, as mentioned previously (pages 21 & 22). It is taken from the curve study mentioned in Chapter 1 and in Chapter 2. The theory to be tested was that of Shortt, who maintained that a car driver entering a curve uses a constant rate of change of centrifugal

TABLE 16. REGRESSION: IMPACT TESTS ON NOTCHED STEEL SPECIMENS

1	2	3	4	5	6
Impact figure -400 y	152	159	202	228	
	136	154	184	214	
	134	149	147	175	
	130	130	143	173	
	114	114	140	172	
	93	99	134	159	
Notch radius x	0·000	0·005	0·010	0·015	
$S(y)$	759	805	950	1121	$3635 = SS_a y$
$\bar{y}_a$	126·5	134·17	158·33	186·33	
$\bar{y}_a S(y)$	96013·5	108004·2	150416·7	209440·2	$563874·5 = S[\bar{y}_a S(y)]$
$S(xy)$	0	4·025	9·500	16·815	$30·340 = SS_a(xy)$
$S(y^2)$	98101	111074	154274	213199	$576648 = SS_a(y^2)$
$S(x)$	0	0·030	0·060	0·090	$0·180 = SS_a(x)$
$S(x^2)$	0	0·00015	0·00060	0·00135	$0·00210 = SS_a(x^2)$

$S(x) = 0·180 \quad \bar{x} = 0·0075 \quad SS_a(y) = 3635 \quad \bar{y} = 151·458$

$S(x^2) = 0·00210 \qquad SS(y^2) = 576648$
$\bar{x}S(x) = 0·00135 \qquad \bar{y}SS(y) = 550551$
$S(x-\bar{x})^2 = 0·00075 \qquad S(y-\bar{y})^2 = 26097$

$S(xy) = 30·3400 \qquad S[\bar{y}_a S(y)] = 563874·5$
$nxy = 27·2625 \qquad \bar{y}S(y) = 550551·0$
$S[x(y-\bar{y})] = 3·0775 \qquad S[n_a(\bar{y}_a-\bar{y})^2] = 13323·5$

$$b = \frac{S[x(y-\bar{y})]}{S(x-\bar{x})^2} = \frac{3·0775}{0·00075} = 4103·33$$

Equation $Y = 551·5 + 4103·4(x-0·0075)$
whence
$$Y = 520·7 + 4103x$$

Note that $\bar{y} = 151·46 + 400 = 551·5$

$S(y-\bar{y})^2 = 26097$
$bS[x(y-\bar{y})] = 12629$
$S(y-Y)^2 = 13309$

$$s^2 = \frac{13309}{22} = 605·0 \qquad s = 24·6$$

$$t = \frac{4103·4\sqrt{0·00075}}{24·6} = 4·60$$

$N = 22 \quad p < 0·01$: very highly significant

Analysis of variance

Item	Sum of squares		N	Mean square	Ratio
Between arrays	$S[n_a(\bar{y}_a-\bar{y})^2]$	13323	3	638·7	1·84
Within arrays	$SS_a(y-\bar{y}_a)^2$	12774	20		
Total	$S(y-\bar{y})^2$	26097	23		
Variation between arrays due to linear regression	$b\,S[x(y-\bar{y})]$	12629	1		
Deviation from regression	$S[n_a(\bar{y}_a-Y_a)^2]$	694·5	2	347·2	
Total	$S[n_a(\bar{y}_a-\bar{y})^2]$	13323·5			

$N_1 = 20 \quad N_2 = 2 \quad p > 0·20$: not significant

Check calculation

x	0	0·005	0·010	0·015
Y from eqn.	520·69	541·20	561·72	582·23
$\bar{y}_a$ from table	526·50	534·17	558·33	586·83
$\lvert\bar{y}_a - Y\rvert$	5·81	7·03	3·39	4·60
$(\bar{y}_a - Y)^2$	33·76	49·42	11·49	21·16

$S(\,) = 115·83$

$S[n_a(\bar{y}_a - Y)^2] = 6 \times 115·83 = 694·5$
$n_a = 6$

acceleration of 1 ft/sec^3. In practical terms, this means that he applies a constant rate of change of force at the steering wheel. As mentioned above, this force was measured with a recording accelerometer. It became clear early in the work that the driver did not use a constant rate of change of force, and therefore that Shortt's theory was not justified. The figures seemed to show that there was some tendency for drivers to turn the wheel faster when entering a sharp curve than when entering a flat one. In other words, the less the radius of the curve, the greater the rate of change of the force applied at the wheel. The tests were then continued, allowing drivers complete freedom to take curves in any way they fancied, and the force was measured and recorded. The problem then resolved itself into finding the relation between the lateral ratio j and its rate of change Q, measured in feet per (second)^3.

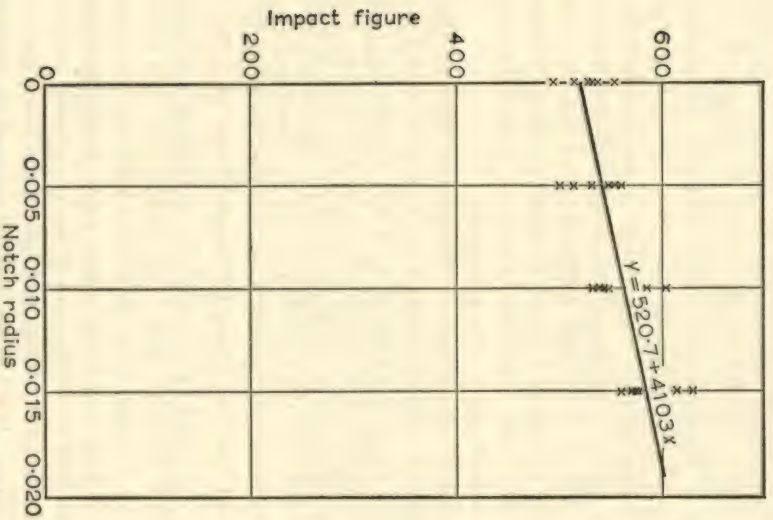


Fig. 11. Impact figures on notch radius for steel specimens, with regression line

The symbol j is now a measure of the lateral ratio, or the force applied at the steering wheel, and does not here have its usual meaning in the rest of the book, as an integer. The lateral ratio j is a pure number. As the curves were taken at random, j varied fairly widely, and naturally Q did the same.

Since the relationship needed was the change of Q with regard to j, and the latter was decided by the radius of the curves recorded, and also since the original theory asserted that the rate of change of force was constant, it is clear that the regression line, if one could be found, was needed in the form $Q = f(j)$, so that Q had to be made the dependent variate y, and j become x. Both the qualities were entered on the cards for sorting, as described above (pages 5 *ff.*) The cards were then first sorted into groups of j, which groups then became the arrays; then these arrays were again sorted into groups of Q, and these were counted, so that now both x and y were grouped.

A rough plot of the means of the groups showed that there was reason to suspect that the regression line would be curved if the equation was obtained in the direct form of $Q = f(j)$. Moreover, Q did not seem to be normally distributed. On these grounds, therefore, it was thought that if Q were expressed in the form of its natural logarithm it would be better distributed, and the equation would be obtained in a straight line relationship between $\log_e Q$ and j, i.e. in the form $\log_e Q = f(j)$. The cards were then resorted, using the same groups of j, but into groups of $\log_e Q$. The resulting figures, and the analysis, are given in Table 17. If the preceding analyses have been understood this table should be clear. It will be noticed once again that some of the figures are self-checking. Thus $S(y)$ and $S(y)^2$ occur twice in the working. The first of these will be seen to be the sum of the fy column, column 17, and it occurs again as the sum of the third line of the lower panel of the main table, the $S(y)$ line. $S(y^2)$ is the sum of column 18, and also the sum of the $S(y^2)$ line, the fourth line from the bottom in the lower panel of the main table. Similarly $S[n_a(y_a-\bar{y})^2]$, which is of great importance in the analysis of variance, can be found in two ways in the table. As before, terms carried to the analysis of variance table are trebly underlined.

The table emphasizes that although the t test shows that the regression coefficient is highly significant, the deviations from regression are not significant, the probability from the analysis being greater than 0·20. It may thus be taken that there is no need to derive any more elaborate equation than that derived at the bottom left of Table 17, which finally results in the expression $Q = 1{\cdot}26e^{3{\cdot}54j}$; so we have arrived at our curved regression line from an analysis of straight line form.

Multiple regression

A further development of regression is when a dependent variate y varies with several other variates, which we may call $x_1, x_2, x_3 \ldots x_j$. For example, it was found eventually in the curve study that the proportion of transition used by drivers appeared to depend on the speed, the radius of the curve, its length, and also powers and cross-products of these three functions. In these circumstances, it is possible to derive what is called a *multiple regression equation.* The coefficients of the independent variates in this type of expression are called *multiple regression coefficients.* It is also sometimes called 'partial regression'.

Again using Y as the symbol for the dependent variate derived from the multiple regression equation, and using $x_1, x_2, x_3 \ldots x_j$ as the independent variates, we need an expression of the general form

$$Y = (y-\bar{y}) + b_1(x_1-\bar{x}_1) + b_2(x_2-\bar{x}_2) + \ldots b_j(x_j-\bar{x}_j) \quad . \quad . \quad . \quad (29)$$

To get the equation we need, we can construct as many equations as necessary, according to the number of independent variates, again by minimizing the sums of squares. This is done by partially differentiating the general expression with regard to each of the multiple regression coefficients in turn. We get a series of equations as follows:

$$b_1 S(x_1-\bar{x}_1)^2 + b_2 S[(x_1-\bar{x}_1)(x_2-\bar{x}_2)] + b_3 S[(x_1-\bar{x}_1)(x_3-\bar{x}_3)] + \ldots = S[(x_1-\bar{x}_1)(y-\bar{y})] \quad . \quad . \quad . \quad . \quad . \quad (30a)$$

$$b_1 S[(x_1-\bar{x}_1)(x_2-\bar{x}_2)] + b_2 S(x_2-\bar{x}_2)^2 + b_3 S[(x_2-\bar{x}_2)(x_3-\bar{x}_3)] + \ldots = S[(x_2-\bar{x}_2)(y-\bar{y})] \quad . \quad . \quad . \quad . \quad . \quad (30b)$$

$$b_1 S[(x_1-\bar{x}_1)(x_3-\bar{x}_3)] + b_2 S[(x_2-\bar{x}_2)(x_3-\bar{x}_3)] + b_3 S(x_3-\bar{x}_3)^2 + \ldots = S[(x_3-\bar{x}_3)(y-\bar{y})] \quad . \quad . \quad . \quad . \quad . \quad (30c)$$

and so on, as far as necessary.

This is obviously very laborious when there are a large number of independent variates, because it is necessary to solve as many equations as there are independent variates to find the multiple regression coefficients. It may, however, be necessary occasionally, and so an outline of the process, with a comparatively simple example, will be given.

The method will be suggested by an examination of the equations 30*a*, 30*b*, etc., which we have to solve to get the multiple regression coefficients b_1, b_2, b_3, etc. It will be seen that the coefficients of the multiple regression coefficients are the various sums of squares and deviation cross-products of the variates. Taking eqn. 30*a* as an example, the coefficient of b_1 in it is the sum of squares of x_1, that of b_2 is the deviation cross-product of x_1 and x_2 and so on. This will be illustrated

TABLE 17. REGRESSION: LATERAL RATIO ON CURVES AGAINST THE LOGARITHM OF ITS RATE OF CHANGE FOR ENTRANCE TRANSITIONS

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
j / $\log_e Q$	0·025 to 0·050	0·05 to 0·075	0·075 to 0·100	0·100 to 0·125	0·125 to 0·150	0·150 to 0·175	0·175 to 0·200	0·200 to 0·225	0·225 to 0·250	0·250 to 0·275	0·275 to 0·300	0·300 to 0·325	0·325 to 0·350	$S(f)$	y	fy	fy^2
2·5 to 2·75	—	—	—	—	—	2	—	—	—	—	2	—	—	4	15	60	900
2·25 to 2·50	—	—	2	1	2	—	1	1	—	—	2	—	2	11	14	154	2156
2·0 to 2·25	—	1	2	4	—	2	2	2	3	3	3	6	—	28	13	364	4732
1·75 to 2·0	—	2	9	7	4	8	10	4	4	4	1	4	1	58	12	696	8352
1·50 to 1·75	2	10	8	20	20	19	8	10	12	4	9	3	6	131	11	1441	15851
1·25 to 1·50	3	8	18	22	14	20	7	11	15	20	9	9	2	158	10	1580	15800
1·00 to 1·25	3	21	17	37	16	25	15	22	12	8	10	9	2	197	9	1773	15957
0·75 to 1·00	5	31	17	43	16	28	20	15	16	6	5	3	—	205	8	1640	13120
0·50 to 0·75	8	40	30	43	27	33	16	17	5	7	3	2	3	224	7	1638	11466
0·25 to 0·50	5	33	39	39	22	26	10	18	9	7	1	2	—	211	6	1266	7596
0·0 to 0·25	5	36	22	33	27	16	6	11	1	5	—	—	—	162	5	810	4050
−0·25 to 0·0	6	16	17	26	6	8	5	1	2	3	—	—	—	90	4	360	1440
−0·5 to −0·25	4	12	11	6	2	6	—	2	1	—	—	—	—	44	3	132	396
−0·75 to −0·5	5	13	16	8	5	5	1	—	—	—	—	—	—	53	2	106	212
−1·0 to −0·75	1	5	3	5	1	1	1	—	—	—	—	—	—	17	1	17	17
−1·25 to −1·00	1	3	1	1	—	—	—	—	—	—	—	—	—	6	0	0	0
$n_a = S(f)$	48	231	212	295	162	199	102	114	80	67	45	38	16	1609		12037	102045
x	0	1	2	3	4	5	6	7	8	9	10	11	12	$S(\ )$			
$S(y)$	277	1453	1409	2110	1205	1538	838	918	710	579	455	380	165	12037			
$\bar{y}_a$	5·771	6·290	6·646	7·153	7·438	7·729	8·216	8·053	8·875	8·642	10·111	10·000	10·313	7·4810		$\bar{y}_a S(y) = 90049{\cdot}4$	
$\bar{y}_a S_a(y)$	1598·5	9139·4	9364·5	15091·9	8963·1	11886·6	6884·7	7392·3	6301·2	5003·6	4600·6	3800·0	1701·5	91727·9			
$S_a(x)$	0	231	424	885	648	995	612	798	640	603	450	418	192	$S(x) = 6896$			
$S_a(x^2)$	0	231	848	2655	2592	4975	3672	5586	5120	5427	4500	4598	2304	$S(x^2) = 42508$			
$S(xy)$	0	1453	2818	6330	4820	7690	5028	6426	5680	5211	4550	4180	1980	$S(xy) = 56166$			
$S(y^2)$	1955	10621	11121	17092	10033	13212	7516	7952	6666	5367	4795	3944	1771	$S(y^2) = 102045$			
$(\bar{y}_a - \bar{y})$	−1·710	−1·191	−0·835	−0·328	−0·043	0·248	0·735	0·572	1·394	1·161	2·630	2·519	2·832				
$n_a(\bar{y}_a - \bar{y})^2$	140·3	327·6	147·8	31·7	0·3	12·2	55·1	37·3	155·4	90·3	311·3	241·1	128·3	$S(\ ) = 1678{\cdot}7 = S[n_a(\bar{y}_a - \bar{y})^2]$			
$S(y - \bar{y})^2$	356·5	1481·6	1756·5	2000·1	1069·9	1325·4	631·3	559·7	364·8	363·4	194·4	144·0	69·5	$S(\ ) = 10317{\cdot}1 = SS_a(y - \bar{y}_a)^2$			

$S[n_a(\bar{y}_a - \bar{y})^2] = 91727{\cdot}9 - 90049{\cdot}4 = 1678{\cdot}5$

$S(Y - \bar{y})^2 = bS[y(x - \bar{x})] = 0{\cdot}3534 \times 4576{\cdot}6 = 1617{\cdot}4$

Analysis of Variance

		Sum of squares	Degrees of freedom	Mean square	Ratio
Between arrays	$S[n_a(\bar{y}_a - \bar{y})^2]$	1678·5	12	6·46	1·16
Within arrays	$SS_a(y - \bar{y}_a)^2$	10317·1	1597		
	$S(y - \bar{y})^2$	11995·6	1609		
Linear regression	$S(Y - \bar{y})^2$	1617·4	1	5·55	
Deviation from regression	$S[n_a(\bar{y}_a - Y_a)^2]$	61·1	11		
	$S[n_a(\bar{y}_a - \bar{y})^2]$	1,678·5	12		

Variance ratio for $N_1 = \infty$ and $N_2 = 11$ of 1·16, $p > 0{\cdot}20$: not significant.

Formation of equation: $\bar{x} = 4{\cdot}2858 \quad \bar{y} = 7{\cdot}481 \quad b = 0{\cdot}3534$

$$Y = a + b(x - \bar{x}) = \bar{y} + b(x - \bar{x}) = 7{\cdot}481 + 0{\cdot}3534(x - 4{\cdot}286)$$
$$= 7{\cdot}481 + 0{\cdot}3534x - 1{\cdot}515 = 5{\cdot}966 + 0{\cdot}3534x = Y$$

but $Y = 4 \log_e Q + 4{\cdot}5$ and $x = 40j - 1{\cdot}5$ } converting working units.

Substituting, $\log_e Q = 0{\cdot}2334 + 3{\cdot}54j$ and $Q = 1{\cdot}26e^{3{\cdot}54j}$

$S(x) = 6896 \qquad \bar{x} = 4{\cdot}2858$

$S(x^2) = 42508 \qquad S(x - \bar{x})^2 = 12952{\cdot}5$

$S(y) = 12037 \qquad \bar{y} = 7{\cdot}4810$

$S(y^2) = 102045 \qquad S(y - \bar{y})^2 = 11995{\cdot}6$

$S(xy) = 56166$

$n\bar{x}\bar{y} = 51589{\cdot}3$

$S[y(x - \bar{x})] = 4576{\cdot}7$

$$b = \frac{S[y(x - \bar{x})]}{S(x - \bar{x})^2} = \frac{4576{\cdot}7}{12952{\cdot}5} = 0{\cdot}3534$$

$S(y - Y)^2 = 10378$

$$s^2 = \frac{1}{1607} \times 10378 = 6{\cdot}458$$

$s = 2{\cdot}542$

$$t = \frac{0{\cdot}3534\sqrt{12953}}{2{\cdot}542} = 15{\cdot}82$$

$p < 0{\cdot}001$: highly significant

in the example, which is given in Table 18. In this table the figures are simple, to reduce the arithmetic, but it serves to illustrate the method, which would be similar for more variates, and with grouped data, though in the latter case the arithmetic might be very heavy.

Those who have studied road safety matters will be familiar with Smeed's equation predicting the number of road deaths per year, denoted by D, in any country, related to the number of registered motor vehicles in it, N, and the population P. This equation is

$$D = 0{\cdot}0003(NP^2)^{1/3}$$

It expresses the average number of deaths per year with reasonable accuracy, and it has done so for a number of years, but there have been some fairly large variations. It is thus possible that some other factors affect the figures. I have worked out a corresponding equation for Europe, but including the length of roads in the country, in kilometres, as well. This gives three independent variates, the population, the number of vehicles, and the length of the roads. The figures were taken from the United Nations publication shown in the table.

The analysis is set out in Table 18, in which the number of deaths is naturally the dependent variate y; and the population, the number of motor vehicles, and the length of roads are respectively x_1, x_2 and x_3. In the table, columns 1 to 5 give the figures, though, again to save arithmetic, they are divided by various powers of 10, as shown at the head of the column. The figures are in any case not known more accurately than is shown. Next, in columns 6 to 15, the various relationships between the variates are calculated, to a fairly high degree of accuracy, because, as usual, we have the differences of fairly large numbers. The sums of squares and deviation cross-products are then worked out in the usual manner in the lower lines of the table. These are then substituted in the equations at the bottom left of the table. Their solution follows usual practice, and is not shown in the table, but the multiple regression coefficients resulting from the solution are given underneath the equations. Below that again, the regression equation is constructed by the substitution in the general expression, eqn. 29, of the values of $\bar{y}$, and of the means of the variates found in the main table, as well as the coefficients. Turning this into the symbols for the various functions, we then have derived the equation

$$d = 0{\cdot}0303p + 0{\cdot}1185m + 0{\cdot}0760l - 0{\cdot}327$$

In this, d is the expected number of road accident deaths in the country per year, in thousands; p is the population of the country in millions; m is the number of registered motor vehicles in hundreds of thousands;

TABLE 18. MULTIPLE REGRESSION: ROAD DEATHS AGAINST POPULATION, OF MOTOR VEHICLES AND LENGTH OF ROADS FOR COUNTRIES IN EUROPE IN 1955*

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Country	Deaths in 1955 y	Population x_1	No. of motor vehicles x_2	Length of roads, km x_3	y^2	x_1y	x_2y	x_3y	x_1^2	x_1x_2	x_1x_3	x_2^2	x_2x_3	x_3^2
	$\times 10^2$	$\times 10^4$	$\times 10^4$	$\times 10^3$										
Belgium	0·78	8·82	5·75	0·92	0·608	6·879	4·485	0·718	77·792	50·715	8·114	33·062	5·290	0·846
Denmark	0·58	4·41	4·44	0·58	0·336	2·558	2·576	0·336	19·448	19·580	2·558	19·714	2·575	0·336
France	7·55	43·27	43·82	7·20	57·003	326·689	330·841	54·360	1872·193	1896·091	311·544	1920·937	315·504	51·840
W. Germany	11·07	49·77	58·63	2·48	122·545	550·954	649·034	27·454	2477·052	2918·015	123·429	3437·477	145·402	6·150
Gt. Britain	5·53	49·57	59·65	3·01	30·581	274·122	329·864	16·645	2457·184	2956·850	149·206	3558·123	179·546	9·060
Italy	5·37	47·16	39·31	1·97	28·837	253·249	211·095	10·579	2224·066	1853·860	92·905	1545·276	77·440	3·801
Luxembourg	0·05	0·30	3·72	0·25	0·003	0·015	0·186	0·013	0·090	1·116	0·075	13·838	0·930	0·063
Netherlands	1·49	10·57	10·30	3·85	2·220	15·749	15·347	5·737	111·725	108·871	40·695	106·090	39·655	14·823
Norway	0·21	3·39	2·06	4·74	0·044	0·712	0·433	0·995	11·492	6·983	16·069	4·244	9·764	22·468
Sweden	0·87	7·26	10·47	1·47	0·757	6·316	9·110	1·279	52·717	76·012	10·672	109·621	15·391	2·161
Switzerland	0·93	4·93	5·40	0·50	0·865	4·584	5·022	0·465	24·305	26·622	2·465	29·160	2·700	0·250
Turkey	0·98	20·95	0·85	4·72	0·960	20·530	0·833	4·626	438·902	17·807	98·884	0·723	4·013	22·278
Yugoslavia	0·50	17·27	0·49	8·16	0·250	8·634	0·245	4·080	298·253	8·462	140·923	0·240	3·998	66·586
$S(\ )$	35·91	267·67	244·89	39·85	245·009	1470·991	1559·071	127·287	10065·219	9940·984	997·539	10778·505	802·208	200·742
Means etc.	2·762	20·59	18·84	3·066	99·194	739·387	676·462	110·078	5511·324	5042·283	820·511	4613·163	750·682	122·156
$n = 13$					145·815	731·604	882·609	17·209	4553·895	4898·701	177·028	6165·342	51·526	78·586
					$= S(y-\bar{y})^2$	$S[x_1(y-\bar{y})]$	$S[x_2(y-\bar{y})]$	$S[x_3(y-\bar{y})]$	$S(x_1-\bar{x}_1)^2$	$S[x_1(x_2-\bar{x}_2)]$	$S[x_1(x_3-\bar{x}_3)]$	$S(x_2-\bar{x}_2)^2$	$S[x_2(x_3-\bar{x}_3)]$	$S(x_3-\bar{x}_3)^2$

* Figures from United Nations Publication 1957 II. E/Mim. 7: Statistics of Road Traffic Accidents in Europe (except for Great Britain, which is from *Basic Road Statistics*)

(*Continued overleaf*)

TABLE 18 (*Continued*)

Equations

$$4553{\cdot}895b_1 + 4898{\cdot}701b_2 + 177{\cdot}028b_3 = 731{\cdot}604$$
$$4898{\cdot}701b_1 + 6165{\cdot}342b_2 + 51{\cdot}526b_3 = 882{\cdot}609$$
$$177{\cdot}028b_1 + 51{\cdot}526b_2 + 75{\cdot}586b_3 = 17{\cdot}209$$

from solution of these equations (not included here)

$b_1 = 0{\cdot}030331 \quad b_2 = 0{\cdot}118469 \quad b_3 = 0{\cdot}076048$

Regression equation

$$(Y-\bar{y}) = b_1(x_1-\bar{x}_1)+b_2(x_2-\bar{x}_2)+b_3(x_3-\bar{x}_3)$$
$$Y-2{\cdot}762 = 0{\cdot}0303(x_1-20{\cdot}59)+0{\cdot}1185(x_2-18{\cdot}84)+0{\cdot}0760(x_3-3{\cdot}066)$$

whence, by simplification and clearing terms.

$$Y = 0{\cdot}0303x_1+0{\cdot}1185x_2+0{\cdot}0760x_3-0{\cdot}327$$

$$S(y-Y)^2 = S(y-\bar{y})^2-\{b_1S[x_1(y-\bar{y})]+b_2S[x_2(y-\bar{y})]+b_3S[x_3(y-\bar{y})]\}$$
$$= 145{\cdot}815-\{22{\cdot}190+104{\cdot}552+1{\cdot}309\}$$
$$= 145{\cdot}815-128{\cdot}051 = 17{\cdot}764$$

Analysis of variance

Item	N	Sum of squares	Mean square	Ratio
Regression: $S\{bS[x(y-\bar{y})]\}$	3	128·05	42·7	21·6
Deviation from regression: $S(y-Y)^2$	9	17·764	1·98	
Total	12	145·814		

$p < 0{\cdot}01$: highly significant

Check calculation for $S(y-Y)^2$

Country	y	Y	$(y-Y)$	$(y-Y)^2$
Belgium	0·78	0·69	+0·09	0·01
Denmark	0·58	0·38	+0·20	0·04
France	7·55	6·72	+0·83	0·69
W. Germany	11·07	8·32	+2·75	7·56
Gt. Britain	5·53	8·74	−2·94	8·64
Italy	5·37	5·91	−0·54	0·29
Netherlands	1·49	1·51	−0·02	0·00
Luxembourg	0·05	0·14	−0·09	0·01
Norway	0·21	0·38	−0·17	0·03
Sweden	0·87	1·25	−0·38	0·14
Switzerland	0·93	0·50	+0·43	0·19
Turkey	0·98	0·77	+0·21	0·04
Yugoslavia	0·50	0·87	−0·37	0·14
$S(\ )$				17·78

cf. 17·76 in table

and l is the length of its roads in hundreds of thousands of kilometres. To simplify the arithmetic, as mentioned above, the original figures have been divided by these powers of ten to reduce the arithmetic.

On p. 84 to the left of the table, $S(y-Y)^2$ is derived from the expression

$$S(y-Y)^2 = S(y-\bar{y})^2-\{b_1S[x_1(y-\bar{y})]+ +b_2S[x_2(y-\bar{y})]+b_3S[x_3(y-\bar{y})]\} \quad (31)$$

$S(y-Y)^2$ is the sum of squares of the differences between the values of y derived from the multiple regression equation, and those actually observed. That is, it is the sum of squares of the deviations from regression. The first term on the right-hand side of eqn. 31 is the main sum of squares, i.e. that corresponding to that of the grand array. The second term is the sum of the multiple regression coefficients, each multiplied by the deviation cross-product of y and its own variate; that is, for example, b_2 is multiplied by the deviation cross-product of y and x_2, and so on. This term can, of course, be extended indefinitely according to the number of independent variates. In this case there are three terms, and they are contained within the brace bracket. Their sum is the sum of squares removed by the regression function. Once again, the deviation sum of squares is derived directly in a separate table as a check on the arithmetic, and the agreement is reasonably good. Here this table also serves another purpose, given below.

The analysis of variance is done as usual, and its object is to test whether the sum of squares of the deviations from regression is in the same proportion to that removed by the regression function as would be expected from random sampling. The number of degrees of freedom of the sum of squares removed by the regression function is the same as the number of independent variates, in this case three. This leaves nine for the deviation degrees of freedom. Here the probability of getting such a variance ratio is less than 0·01, and so the deviations from regression are very definitely not significant. The sum of squares removed by deviations from regression is very much smaller than that removed by the regression function.

The column for $(y-Y)$ in the check calculation shows that the two countries which differ most widely from regression are West Germany, which is about 2750 deaths above the number predicted from the equation, and Great Britain, which is about 2940 below it, both, of course, on the figures for 1955. Fig. 12 shows the actual deaths plotted against the number predicted from the equation. This is a very usual form of graph for this type of information. The figures for plotting it are derived from the check calculation table. If the actual figure for

any country was exactly the same as that predicted the plotted point would coincide with the diagonal line. In Fig. 12 the countries are denoted by their international registration letters. The figures for Great Britain are also plotted for the years 1957, 1958, and 1959, the figures being again taken from *Basic Road Statistics*.* The figure for Great

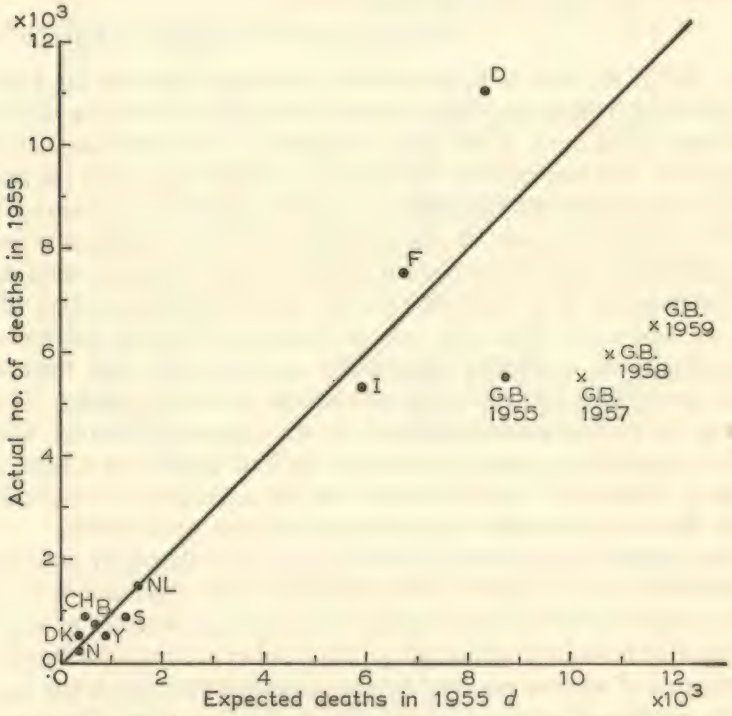


Fig. 12. Deaths in road accidents for Europe in 1955. Actual numbers plotted against those calculated from the regression equation

Values of d are calculated from the equation $d = 0{\cdot}0303p + 0{\cdot}1185m + 0{\cdot}076l - 0{\cdot}327$.

× Later figures for Great Britain not used in regression table.

Britain continues to be well below predicted, in spite of the sharp upward rise shown in Fig. 10, though admittedly it does not necessarily follow that the equation applies for the succeeding years.

It is not essential that x_1, x_2, x_3, etc., should be different variates. They could equally well be functions of the same variate, such as x^2, x^3, $\log x$, or what you will. This could be used to obtain a curved regression line, by giving an equation of the form

$$y = a + bx + cx^2 + dx^3 + \ldots$$

* Issued annually by the British Road Federation.

The fitting of higher powers can, however, be done in another way in stages, and the method is described in various textbooks, such as those by Mather[4] and Fisher.[6] On the other hand, the addition to the table as set out here presents little more difficulty than the addition of extra columns, so that part of the work may not be much greater, but, of course, the solution of the equations becomes more and more laborious as they increase in number.

The variates x_1, x_2, etc., could also be cross-products, or other functions of two or more variates, and so we could obtain a regression equation of the form

$$y = a + bx_1 + cx_2 + dx_1 x_2 + \ldots$$

The number of different possibilities is almost infinite.

Example 10.1. Extending the analysis of Examples 4.1, 5.1 and 8.1, find whether significant regression equations exist for the death rates in the two states of Connecticut and Rhode Island. Plot the annual death rates on a dot diagram, and add on the regression lines, if any.

CHAPTER 11

Correlation

If, as in Chapter 10, we have two variates which may or may not be related to each other in some way, we may use another, and slightly shorter, test to find whether they are indeed related. This is finding the *correlation coefficient*, which is usually denoted by the letter r. It is a number which may vary between $+1$ and -1. If the variates are x and y, we have seen in the last chapter that we can find the regression coefficient of x on y, or y on x, at choice. The correlation coefficient is the geometric mean of these two functions. It is a measure of the angle between the two regression lines. If it is $+1$ the two lines are coincident, and sloping upwards, and the two variates are fully related. Similarly, if it is -1 the two lines coincide and slope downwards, and again the two variates are fully related. If r is zero, the two lines are at right-angles, and the variates are not related.

The correlation coefficient is given by the expression

$$r = \frac{S[y(x-\bar{x})]}{\sqrt{[S(x-\bar{x})^2 . S(y-\bar{y})^2]}} \quad . \quad . \quad . \quad . \quad (32)$$

that is, it is the deviation cross-product divided by the square root of the product of the two sums of squares. Also, if the regression coefficient of x on y is b_x and that of y on x is b_y, we have

$$r = \sqrt{(b_x b_y)} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (33)$$

The significance of r may be tested by means of a t test with

$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (34)$$

The number of degrees of freedom of this t is $(n-2)$.

As an example, we may use Table 17 (page 80) and find the cor-

relation between $\log_e Q$ and j. From the table, the deviation cross-product is 4576·7, and the two sums of squares are 12952·5 and 11995·6. Thus

$$r = \frac{4576{\cdot}5}{\sqrt{(12952{\cdot}5 \times 11995{\cdot}6)}} = 0{\cdot}367$$

Then

$$t = \frac{0{\cdot}367 \times \sqrt{1607}}{\sqrt{(1-0{\cdot}367^2)}} = 15{\cdot}9$$

and so the coefficient is very highly significant.

The correlation coefficient merely tells us that the variates are connected, but it does not tell us how they are connected. It therefore does not give us as much information as regression, and there is some tendency for its use to be dropping out, and so it is only briefly mentioned here. It could be used to save work in the table of regression, since by first working out the deviation cross-product and the two sums of squares, r could be found, and if this was significant it would then indicate that the work of continuing with the regression table was worth while.

CHAPTER 12

The Chi-Square Test for Goodness of Fit

The χ^2 test compares observations with some known or suspected hypothesis, and the appropriate expression for χ^2 has been given in eqn. 11. For convenience this will be repeated. It is

$$\chi^2 = S\left[\frac{(O-E)^2}{E}\right] \quad . \quad . \quad . \quad . \quad (11) \text{ repeated}$$

It is most important to note that this form of the expression for χ^2 *may only be used for frequencies, and not for any form of measurement, average, or percentage.* As the first and simplest example of its use, figures taken from a paper by me will be used. They also illustrate the use of the binomial distribution. We have seen above (page 18) that if two dice are thrown, the probability of getting two sixes, or one six, or no sixes at all, is given by the expansion of $(p+q)^2$, and that this results in the expectation that if we make thirty-six throws, we expect to get two sixes once, one six ten times, and no sixes at all twenty-five times. But χ^2 is not accurate for an expectation of less than six, and so to apply the test with any confidence the dice must be thrown at least $6 \times 36 = 216$ times. This was done with a pair of dice, and the result is tabulated in Table 19, with the expectations also shown, and also the values of χ^2 worked out from eqn. 11.

We have seen that the number of degrees of freedom is the number of independent comparisons which can be made in the data. Applying this to the present problem, we have first that the total number of throws is entirely at the choice of the experimenter, which fixes the total at the foot of the table. We also have that the E column, column 3, is fixed by the hypothesis as worked out above, and it cannot be varied. Then in the left-hand column, column 2, only two of the figures can be filled in arbitrarily, and after this has been done, the third must

follow by subtraction of the first two from the total already fixed. So we see that the table has two degrees of freedom.

In filling in column 4, that for χ^2, the values are calculated from eqn. 11. Taking the first line, O is 1, and E is 6; so $|O-E| = 5$, and $\chi^2 = 5^2/6 = 25/6 = 4{\cdot}167$, and so on. From Table H we see that the probability of getting a value of χ^2 of 5·57 for $N = 2$ is between 0·05 and 0·1, rather nearer the former, so that it is very nearly significant at the 0·05 level, and the test can be taken as a fairly strong indication that the dice are biased.

TABLE 19. THE χ^2 TEST: BIAS OF DICE

1	2	3	4
	O	E	χ^2
Two sixes	1	6	4·167
One six	54	60	0·600
No sixes	161	150	0·807
Total	216	216	5·574

The use to be made of this information would depend on the importance attached to the bias. The dice used for the test were intended for a children's game, and so the bias is unimportant. If they were to be used for gambling, we should be more circumspect. This would be more especially so because we notice from the table that the principal deficiency from expectation is a shortage of two sixes!

There are many forms of the equation for χ^2, a few of which are mentioned here. One is

$$\chi^2 = S[(a-pn)^2/pn] \quad . \quad . \quad . \quad . \quad . \quad . \quad (35)$$

in which n is as usual the number of observations, p is the expected proportion, and a the observed number. Another form of this is

$$\chi^2 = S\left(\frac{a^2}{pn}\right) - n \quad . \quad . \quad . \quad . \quad . \quad . \quad (36)$$

A frequent use of χ^2 in traffic studies is to find the significance of some change in a factor before and after some alteration in conditions has been made, making a comparison with some other place in which no alteration had taken place, which will be called the control. Let the number of factors before the alteration had taken place be a; the number after the alteration be b; and the corresponding number of factors for the control, respectively before and after the alterations had

been made at the test place, and for the same period, be c and d. We can then set up a table in the form shown in Table 20.

TABLE 20. THE χ^2 TEST: BEFORE-AND-AFTER TESTING AT A SINGLE SITE

	Before	After	$S(\ )$
Test	a	b	$(a+b)$
Control	c	d	$(c+d)$
$S(\ )$	$a+c$	$b+d$	$n=(a+b+c+d)$

We can then calculate χ^2 from the expression, which is derived from eqn. 11:

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad . \quad . \quad . \quad . \quad (37)$$

This expression involves the calculation of the numerator to a high degree of accuracy if the figures are at all large, because it is the square of the difference between two large quantities; although once the square has been done, the rest can be done to a lower degree of accuracy, since χ^2 is only needed to two or three places. There is, however, another form of the expression for χ^2 which can be done wholly on the slide rule. It is due to Wolf. In this we calculate

$$\frac{a}{(a+c)} - \frac{b}{(b+d)} = A \quad . \quad . \quad . \quad . \quad . \quad . \quad (38a)$$

$$\frac{a}{(a+b)} - \frac{c}{(c+d)} = B \quad . \quad . \quad . \quad . \quad . \quad . \quad (38b)$$

and then

$$\chi^2 = nAB \quad . \quad . \quad . \quad . \quad . \quad (38c)$$

However it is calculated, the table has one degree of freedom, because only one of the factors a, b, c or d can be written in arbitrarily, and then the rest must follow from the totals.

As an example of this, the accidents on a length of road on which a speed limit was imposed by reason of the erection of a system of street lighting by a parish council may be used. The limit had previously been refused by the highway authority. The object is to find out whether the number of accidents for a three year period before, and for a similar period after, the limit was imposed, were significantly different. For the control, the number of accidents in the rest of the county was used.

The figures are set out in Table 21, using the notation of Table 20. Thus in the three years before the imposition of the limit there were

TABLE 21. THE χ^2 TEST: ACCIDENTS BEFORE AND AFTER THE IMPOSITION OF A SPEED LIMIT

	Before	After	$S(\ )$
Test	$a = 12$	$b = 20$	$a+b = 32$
Control	$c = 4339$	$d = 5340$	$c+d = 9679$
$S(\ )$	$a+c = 4351$	$b+d = 5360$	$n = 9711$

$$\left.\begin{aligned} \frac{a}{(a+c)} &= \frac{12}{4351} = 0.00276 \\ \frac{b}{(b+d)} &= \frac{20}{5360} = \underline{0{\cdot}00373} \\ A &= -0.00097 \end{aligned}\right\} \text{ from eqn. 38}a$$

$$\left.\begin{aligned} \frac{a}{(a+b)} &= \frac{12}{32} = 0.375 \\ \frac{c}{(c+d)} &= \frac{4339}{9679} = \underline{0{\cdot}448} \\ B &= -0{\cdot}073 \end{aligned}\right\} \text{ from eqn. 38}b$$

Then $\chi^2 = 9711 \times 0{\cdot}00097 \times 0{\cdot}073 = 0{\cdot}69$ from eqn. 38c

$0{\cdot}5 > p > 0{\cdot}1$: not significant

12 accidents, and there were 20 in the three years after, while the corresponding numbers in the rest of the county were respectively 4339 before and 5340 after.

It is useful to note, and to remember, that if χ^2 is smaller than the number of its degrees of freedom it is certainly not significant. It is here less than one, and in looking it up in the table we see that then the probability for one degree of freedom is between 0·5 and 0·1 (nearer the former); so it is clearly not significant. We therefore cannot conclude from the figures that the speed limit made any difference to the accident figures on this stretch of road, when allowance has been made for the accidents in the rest of the county.

We can also solve more complex problems by the use of χ^2. Continuing with the question of speed limits, it might be urged against any conclusion from the last analysis that while the speed limit may not have reduced the number of accidents, it would still reduce their severity. The figures from the stretch of road previously considered are too small to enable this test to be done, but figures are available for

the accidents in one county for two years before and two years after the imposition of the general 30 miles/h speed limit in 1935, taken for the stretches of road on which the limit was imposed. They are set out in Table 22. Here we have two questions to ask. The first is, do the

TABLE 22. ACCIDENTS FOR TWO YEARS BEFORE AND TWO YEARS AFTER THE IMPOSITION OF THE 30 MILES/H SPEED LIMIT IN 1935

Type of accident	Before limit	After limit	Total
Fatal	11	8	19
Injury	280	296	576
Non-injury	371	337	708
Totals	662	641	1303

figures show that the number of accidents were reduced by the limit? The second is, do the figures show that the limit reduced the severity of the accidents?

This kind of table is called a *contingency table*, and if it has i rows and j columns it is called an $i \times j$ contingency table, so that Table 22 is a 3×2 contingency table.

In Table 22 there are three degrees of freedom, because the totals are independent, having been taken from observed figures, and not at the experimenter's choice, as were those of Table 19. The first step is to test whether the two column totals differ significantly. That is to say, we test whether the speed limit can be thought of as having reduced the number of accidents, though the hypothesis we actually test is that the limit has had no effect. If this were true, we would expect both totals to be the same. The expression for χ^2 to cover this case is

$$\chi^2 = \frac{(a_1 - a_2)^2}{n} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (39)$$

in which a_1 and a_2 are the numbers in each class, and $n = a_1 + a_2$.

Applying this to the totals in Table 22, we get that

$$\chi^2 = (662-641)^2/1303 = 21^2/1303 = 0{\cdot}338$$

This has one degree of freedom and is clearly not significant, as the value is much smaller than its degree of freedom. The probability is actually between 0·5 and 0·7, and so the figures show no reason whatever to doubt the proposition, which is that the speed limit has not affected the number of accidents.

We may next go on to examine the alternative proposition that the limit has not affected the severity of the accidents, and for this we now take the totals in Table 22 as fixed. On the hypothesis, we would expect that of the total of accidents before the imposition of the limit, 662, the proportion of fatal accidents would be 19/1303, and so the expected number of fatal accidents would be $(19 \times 662)/1303 = 9{\cdot}65$. Similarly, the expected number of injury accidents would be $(576 \times 662)/1303 = 292{\cdot}65$. We then set up a new table, Table 22A, in which we show the expected values in brackets alongside the corresponding observed ones. It is only necessary to work out two of the expected numbers, when the rest can be done by subtraction, because the totals must be equal.

TABLE 22A. THE χ^2 TEST: ANALYSIS OF TABLE 22

Type of accident	Before limit	After limit	Total
Fatal	11 (9·65)	8 (9·35)	19
Injury	280 (292·65)	296 (283·35)	576
Non-injury	371 (359·70)	337 (348·30)	708
Totals	662	641	1303

We now work out another value of χ^2, based on eqn. 11. It will be

$$S\left[\frac{(O-E)^2}{E}\right] = \frac{(11-9{\cdot}65)^2}{9{\cdot}65} + \frac{(8-9{\cdot}35)^2}{9{\cdot}35} + \frac{(280-292{\cdot}65)^2}{292{\cdot}65} + \frac{(296-283{\cdot}35)^2}{283{\cdot}35} + \frac{(371-359{\cdot}70)^2}{359{\cdot}70} + \frac{(337-348{\cdot}30)^2}{348{\cdot}30}$$

$$= 2{\cdot}22$$

In this case there are two degrees of freedom, because the totals are now fixed, and these added to the one we have already used in testing the totals make up the three for the table. For two degrees of freedom the probability of getting this value of χ^2 is between 0·1 and 0·5, and a more detailed table shows that it is in fact slightly greater than 0·3. Here again, the figures show no reason to doubt the proposition that the speed limit did not affect the severity of the accidents either.

In this case no control was used. None is, in fact available, but reference to *Basic Road Statistics** shows that it is doubtful whether the changes in accidents about the time of the imposition of the general speed limit were large enough to affect the issue of the present analysis.

* See footnote on p. 86.

The number of possible uses of χ^2 is very large and other methods of use will be found in various textbooks. One more method will be given here, which is very useful for 'before-and-after' studies. In these, the accidents in a number of places before and after some change or other has been made in road conditions are compared to see whether the change has affected the accidents. This has already been done in Table 21 for one single site, but it does not give us a conclusion in which we can place much confidence, partly because we can never be sure that some extraneous circumstance, such as a local change in traffic conditions, may not have affected the result, and partly because the number of accidents at an isolated place is usually fairly small. This aspect of the matter has already been discussed (page 46) in considering the t test, and so it is not necessary to pursue it further.

If, however, we can compare a number of similar changes, and moreover if we can compare them with a control, such as the number of accidents in a control area (e.g. a whole county), the result will give a much better indication of the effect of the type of change, and we can have much more confidence in any conclusion drawn. The control area should not, of course, have been affected by the change being investigated, or, if it includes it, should be large enough for the change not to matter.

The following method is due to Tanner. Its derivation involves complex mathematics for which the reader is referred to the original paper.*

In the equations which follow, n is the number of sites, b_i the number of accidents at any single site i before the change, and a_i the number of accidents there after the change; B_i and A_i denote the corresponding numbers of accidents for the control, for the same periods as those for which a_i and b_i were taken, and $C_i = A_i/B_i$. The number of accidents at any site for the whole period under review is $n_i = a_i + b_i$. The apparent effect of the change at the site i will be denoted by $k_i = a_i/b_i C_i$. This ratio k_i is the ratio of the accidents after the change to the number which would be expected if the change had no effect and if the accidents there changed in the same way as those for the control. It should be particularly noted that $S(n_i)$ is not the same as n.

The first step is to find k, the average of the values of k_i for all the sites, which is done by means of the expression

$$S\left[\frac{n_i}{(1+kC_i)}\right] = S(b_i) \quad . \quad . \quad . \quad . \quad . \quad . \quad (40)$$

This is solved by trial and error, by summing up the values of $n_i/(1+kC_i)$ for all sites, assuming values of k until the summation

* *Biometrika*, 45 (1958), Parts 3 and 4, pp. 331–342.

equals the sum of all the b_i values for the sites. To find the first value of k to try, work out each value of $b_i C_i$, and add them all up. Then $S(a_i)/S(b_i C_i)$ will give this first value of k, and this will often be found to be very nearly accurate. For testing its significance we test that of $\log_e k$, but for this we must first find χ^2, because its value may affect that of the variance of $\log_e k$. This is done from the expression,

$$\chi^2 = S\left[\frac{(a_i - kb_i C_i)^2}{kn_i C_i}\right] \quad . \quad . \quad . \quad . \quad . \quad (41)$$

This χ^2 has $(n-1)$ degrees of freedom, and it tests whether the effect of the changes has been consistent over the sites. That is, if it proves to be significant, we can assume that the effect of the change has not been the same at all sites.

The estimate of the variance of $\log_e k$, called $\operatorname{var}\log_e k$, is given by the expression

$$\operatorname{var}\log_e k = \frac{(1+\phi)\left[1+\dfrac{2}{S(n_i)}\right]}{S\left[\dfrac{kC_i n_i}{(1+kC_i)^2}\right]} \quad . \quad . \quad . \quad . \quad (42)$$

The value of ϕ in this expression is found from another one

$$\phi = \left[\frac{\chi^2}{(n-1)} - 1\right]\frac{nS(n_i^2)}{S^2(n_i)} \quad . \quad . \quad . \quad . \quad . \quad (43)$$

It is sometimes possible to simplify eqn. 42. If the χ^2 test shows that the results are probably consistent, that is if it gives p greater than 0·2, then $(1+\phi)$ can be taken as unity and omitted from the expression. Then if $S(n_i)$ is reasonably large, say greater than 40, we may also treat $[1+2/S(n_i)]$ as unity. Finally, if most of the values of kC_i are in the range $\frac{1}{2}$ to 2, we may replace the denominator by $\frac{1}{4}S(n_i)$.

The estimate of the standard deviation, the root mean square, of $\log_e k$ will be the square root of $\operatorname{var}\log_e k$, and then dividing $\log_e k$ by this root mean square will then give a value of t with which to test the significance of $\log_e k$ and hence that of k.

The example is a comparison made in August 1959 of the effect on accidents of the realignment and reconstruction of a number of roads in Dorset. These realignments had, however, only been to two-lane standards, and did not include the construction of dual carriageways; although where dual carriageways were to be the final alignment, the standard of the two-lane one had been made to suit them. Strictly

TABLE 23. THE χ^2 TEST: BEFORE-AND-AFTER ANALYSIS OF ACCIDENTS ON RECONSTRUCTED LENGTHS OF ROADS

1	2	3	4	5	6	7	8	9	10	11	12
No. of length	Dates										
	Start	End	b_i	a_i	n_i	B_i	A_i	C_i	$\lvert a_i - kb_iC_i \rvert$	χ^2	$\frac{C_i n_i}{(1+C_i)^2}$
1	5/58	12/58	1	0	1	5621	1430	0·254	0·07	0·07	0·16
2	5/55	10/55	4	2	6	4581	5792	1·264	0·53	0·13	1·48
3	4/55	6/55	2	0	2	4587	5616	1·224	0·71	0·69	0·50
4	1/57	9/57	3	0	3	5340	4935	0·924	0·80	0·80	0·75
5	10/54	3/55	0	1	1	4505	5625	1·248	1·00	2·76	0·25
6	1/57	8/57	2	0	2	5340	4195	0·786	0·46	0·46	0·49
7	4/58	9/58	6	0	6	5602	1901	0·339	0·59	0·59	1·13
8	4/56	3/57	1	0	1	4990	4973	0·997	0·29	0·29	0·25
9	9/57	4/58	0	2	2	5477	2938	0·536	2·00	12·87	0·45
10	7/55	10/55	2	0	2	4634	5792	1·250	0·73	0·73	0·49
11	6/57	10/57	1	0	1	5379	4741	0·881	0·26	0·27	0·25
12	5/58	12/58	2	0	2	5621	1430	0·254	0·15	0·15	0·32
13	9/56	3/58	1	1	2	5211	2938	0·564	0·84	2·16	0·46
14	1/58	7/58	3	0	3	5568	2438	0·438	0·38	0·38	0·64
$S(\ )$			28	6	34					22·25	7·62

$k = 0{\cdot}29$ $\log_e k = -1{\cdot}24$ $\chi^2 = 22{\cdot}25$ $N = 13$ p slightly $> 0{\cdot}05$: just significant

$(1+\phi) = 1{\cdot}94$ var $\log k = 0{\cdot}270$ r.m.s. $\log_e k = 0{\cdot}52$ $t = \frac{1{\cdot}24}{0{\cdot}52} = 2{\cdot}38$ $0{\cdot}05 > p > 0{\cdot}02$: significant

speaking, therefore, the work actually done was not up to full modern standards of visibility. It is sometimes said that as this type of work speeds up the traffic—which is undoubtedly true—it must increase accidents. The logic of this is not immediately apparent, but as the view is often made quite seriously it must be investigated. The figures and working for this investigation are given in Table 23.

In this table, the first column is an identification number for the stretch of road concerned. Column 2 gives the date, taken to the nearest first day of a month, of the start of the work. For this purpose, the start has been taken as the date on which some noticeable change in traffic conditions was made, which might affect accidents. If a new fence had been erected behind a hedge, so that it would not be visible to traffic, this was not taken as a start of work; but if a substantial length of hedge had been taken down, that would be considered as affecting traffic. For length No. 1 something of this sort was done in early May 1958, and so the date in column 2 is given as 5/58, and so on for the rest. Column 3 gives the corresponding date for the end of the work, usually the date when the final surfacing was finished, or sometimes when the 'temporary' surface was finished, if it was of such quality and finish that it could be accepted as adequate for a fairly long period before the final surface was laid. For length No. 1 this was done in December 1958, and so the entry on line 1 for column 2 is 12/58. Column 4 gives b_i, the number of injury accidents on the length before the date given in column 2, for a three year period in all cases. It is best to make the periods three years if possible, though other periods can be taken if necessary. In line 1, then, there was one accident in length 1 between the 1st May 1955 and the 1st May 1958, when work was started. Column 5 shows the number of injury accidents, a_i: either for a three year period after the completion date given in column 3; or if the work had been finished less than three years before 1st August 1959—the date of the analysis—then a_i is the number of accidents during the intervening period, whatever its length. Column 6 is $n_i = (a_i + b_i)$. Columns 7 and 8 are the number of injury accidents in the whole county for the corresponding periods used for b_i and a_i. Thus, for line 1, there were 5621 injury accidents in the whole county during the three year period before 1st May 1958, and there were 1430 in the eight months between 1st December 1958 and 1st August 1959. Then column 9 is $C_i = A_i/B_i$, so that in line 1, $C_i = 1430/5621 = 0{\cdot}254$.

The first nine columns set out the data for the analysis. The initial step in the working is to find k. For this a trial value of k is used. We can derive this from $S(b_iC_i)$, which is 21, and the trial k is $6/21 = 0{\cdot}29$ = the correct value as it happens. Then $n_i/(1+kC_i)$ is then worked

out for each line, and the results added up. The sum should come to $S(b_i)$, which in this case is 28, to satisfy equation 40. The working is not shown in the table, but k comes to 0·29, which means that the number of accidents after the work was done averaged at 29% of what would have been expected if the work had not been done, and if the accidents on those stretches of road had increased in the same proportion as those in the rest of the county. Then $\log_e k = -1{\cdot}24$.

The next step is to find χ^2. To do this, $|a_i - kb_i C_i|$ has been worked out for each stretch, and entered into column 10. In column 11 the term in eqn. 41 is worked out by squaring the quantity in column 10 and dividing by $kn_i C_i$. The sum of this column gives χ^2, which is 22·25. For 13 degrees of freedom the 0·05 point is 22·36; so this is just significant, and we may conclude that the figures imply that the effect of the work is not consistent.

To test the significance of k, we first work out $(1+\phi)$ from eqn. 43. From this $\phi = (22{\cdot}25/13-1)(13\times 118/34^2) = (1{\cdot}71-1)1{\cdot}33 = 0{\cdot}94$, so that $(1+\phi) = 1{\cdot}94$. This is substituted in eqn. 42, the terms for the numerator having been already worked out for each line in column 12, although in this we take k as unity because we are testing whether the true value of k is unity; thus we get that

$$\text{var}\log_e k = \frac{1{\cdot}94(1+2/34)}{7{\cdot}62} = \frac{1{\cdot}94\times 1{\cdot}06}{7{\cdot}62} = 0{\cdot}27$$

Therefore the root mean square $= \sqrt{0{\cdot}27} = 0{\cdot}52$, and $t = \log_e k/(\text{r.m.s. } \log_e k) = 1{\cdot}24/0{\cdot}52 = 2{\cdot}38$. For 13 degrees of freedom this gives a probability of between 0·05 and 0·02, and so is significant. We can therefore conclude that the evidence given in the table shows that the works done have reduced accidents to about one-third of what would have been expected if the works had not been done, and if the accidents on the stretches of road concerned had increased at the same rate as those in the rest of the county, although it shows that we cannot be confident that such work will always have this effect. The data we have here, however, give no reason to think that the proposition—that work of this kind which speeds up traffic will also tend to increase accidents—is justified.

If on any stretch of road reconstructed, or included in any analysis of this type, there had been no accidents either before or after, so that $n_i = 0$, we could not include it in the analysis.

Example 12.1. The following figures were given (*Highway Times*, April 1962) for the accidents on road A127 before and after the introduction of the 'Clearway' system of the complete prohibition of stopping on the carriageway. Do these figures

indicate that the 'Clearway' system has (*a*) reduced accidents or (*b*) reduced their severity?

Type of accident	1959 Before	1961 After
Fatal	4	3
Injury	94	86
Non-injury	74	60
Total	172	149

CHAPTER 13

The Poisson Series

It has been mentioned in Chapter 3 that the Poisson series results when, in the binominal distribution, one or other of the two probabilities, p or q, is of the order $1/r$. When this happens, the probabilities of 0, 1, 2, 3, 4, . . . j events occurring are then given by the terms of the series

$$e^{-m}\left[1,\ m,\ \frac{m^2}{2!},\ \frac{m^3}{3!},\ \ldots\frac{m^j}{j!},\ldots\right] \quad . \ . \ . \quad (44)$$

The sum of the terms within the bracket will be e^m, so the sum of the series will be unity, because $e^{-m}.e^m = e^0 = 1$.

In eqn. 44, m is the mean, and the variance is also m. The mean and the standard deviation apply to the Poisson distribution as usefully as to the normal distribution. The Poisson distribution is applicable when the probability of an event occurring is small, but the number of possibilities of its occurrence is large, and so it is useful for some of the problems in traffic. Taking accidents as an example, we know that the probability of an accident occurring at, say, a cross-roads is small. But on a main road the number of vehicles passing is large, so that some accidents may not be unexpected.

It is also found that traffic itself tends to be distributed along the road in accordance with the Poisson distribution, provided it can run freely, and is not held up by traffic lights, or some form of traffic congestion.* This can be of much use in the design of traffic lights, and similar matters.

As an example of the use of the Poisson series, two counts made in Dorset are given in Table 24. The counts were made for another purpose, to find the average speed of traffic over a stretch of road, but

* ADAMS, W. F.: 'Road traffic considered as a random series', *J. Instn. Civil Engrs.* 4 (1936), p. 121.

TABLE 24. THE POISSON SERIES: TRAFFIC COUNT ON LENGTH OF ROAD

1	2	3	4	5	6	7	8	9	10	11	12
No. of vehicles per interval x	Out						In				
	f	fx	fx^2	Probability	E	χ^2	f	fx	fx^2	E	χ^2
10	0	0	0	0·002	0·12		2	20	200	0·43	
9	0	0	0	0·006	0·38		2	18	162	1·03	0·15
8	4	32	256	0·015	0·95	0·06	2	16	128	2·21	
7	1	7	49	0·036	2·27		3	21	147	4·25	
6	4	24	144	0·073	4·60		4	24	144	7·12	1·37
5	9	45	225	0·128	8·06	0·11	12	60	300	10·24	0·30
4	10	40	160	0·187	11·78	0·27	11	44	176	12·27	0·13
3	11	33	99	0·218	13·73	0·54	4	12	36	11·76	5·12
2	12	24	48	0·191	12·03	0·00	15	30	60	8·45	
1	11	11	11	0·110	6·99	1·04	6	6	6	4·05	6·74
0	1	0	0	0·032	2·02		2	0	0	0·97	
$S(\ )$	63	216	992	0·998	62·93	2·02	63	251	1359	62·76	13·81

$n = S(fx) = 216$

$m = \bar{x} = \frac{216}{63} = 3{\cdot}429$

$S(fx^2) = 992$

$\bar{x}\ S(x) = 740{\cdot}6$

$S(x-\bar{x})^2 = 251{\cdot}4$

$s^2 = \frac{251{\cdot}4}{62} = 4{\cdot}055$

$n = S(fx) = 251$

$m = \bar{x} = \frac{251}{63} = 4{\cdot}173$

$S(fx^2) = 1359$

$\bar{x}\ S(x) = 1047$

$S(x-\bar{x})^2 = 312$

$s^2 = \frac{312}{62} = 5{\cdot}032$

Both have $N = 4$

$0{\cdot}8 > p > 0{\cdot}7$: Differences not significant. Consistent with Poisson

$p < 0{\cdot}01$: Differences highly significant. Not consistent with Poisson

they were made in a form suitable for the present analysis. Observers were stationed with synchronized stop watches at each end of the stretch. They noted the exact time of arrival of each vehicle to one-hundredth of a minute, and also its registration number. It was thus possible to count the number of vehicles arriving at the check point in any given interval of time. In the table, this interval was taken as two minutes, and the left half of the table, columns 2 to 7, headed 'Out', shows eastbound traffic; while the right half, columns 8 to 12, headed 'In', shows the westbound traffic. Column 1 shows the number of vehicles passing in the two-minute interval.

Columns 2 and 8 show the frequency with which the number of vehicles shown in column 1 passed during the interval. Thus column 2 shows that in the 'Out' direction, there were nine cases of five vehicles passing during the two-minute interval, while column 8 shows that in the 'In' direction there were twelve such cases. The fx and fx^2 columns have their usual meanings, the number of vehicles per interval, in column 1, being x. These figures are used as before to calculate the mean and the mean square, the estimate of the variance. These should be equal, and it will be seen that in neither case is there very good agreement. This is not altogether unexpected, and the goodness of fit can be tested by means of a χ^2 test.

The individual expectations can be worked out directly, which is actually done in the 'In' table, column 11, but to illustrate the method the probabilities will be calculated in column 5. Dealing first with the 'Out' table, we have already calculated $\bar{x}$, the mean number of vehicles passing in the two-minute interval, as 3·429. From this we first determine the value of e^{-m}. This can be found from Table D, but it may be helpful to readers to give the method of calculation:

$$\log\log e = \bar{1}\cdot 63778$$
$$\log m = 0\cdot 53516$$
Therefore $$\log\log e^m = \overline{0\cdot 17294}$$

The antilog of this gives $\log e^m$, which is 1·4891, and this is subtracted from zero to give $\log e^{-m}$, because this term is the reciprocal of e^m. The process gives $\log e^{-m}$ as $\bar{2}\cdot 5109$, whence $e^{-m} = 0\cdot 0324$.

If p_j = probability of j cars per interval, the table of probabilities is then worked out as follows:

$$p_0 = e^{-m} = 0\cdot 0324$$
$$p_1 = e^{-m}m = p_0 m = 0\cdot 1111$$
$$p_2 = e^{-m}\frac{m^2}{2!} = p_1 m/2 = 0\cdot 1905$$

$$p_3 = e^{-m}\frac{m^3}{3!} = p_2 m/3 = 0\cdot 2177$$
$$p_4 = e^{-m}\frac{m^4}{4!} = p_3 m/4 = 0\cdot 1866$$
$$p_5 = e^{-m}\frac{m^5}{5!} = p_4 m/5 = 0\cdot 1280$$
$$p_6 = e^{-m}\frac{m^6}{6!} = p_5 m/6 = 0\cdot 0731$$
$$p_7 = e^{-m}\frac{m^7}{7!} = p_6 m/7 = 0\cdot 0358$$
$$p_8 = e^{-m}\frac{m^8}{8!} = p_7 m/8 = 0\cdot 0154$$
$$p_9 = e^{-m}\frac{m^9}{9!} = p_8 m/9 = 0\cdot 0059$$
$$p_{10} = e^{-m}\frac{m^{10}}{10!} = p_9 m/10 = 0\cdot 0020$$
$$p_{11} = e^{-m}\frac{m^{11}}{11!} = p_{10} m/11 = 0\cdot 0006$$
$$S(p) = \overline{0\cdot 9991}$$

We see that each of these terms follows from the preceding one by multiplying by m and dividing by the number of the power of m in the term needed, since

$$\frac{p_j}{j!} = \frac{p \times p^{(j-1)}}{j \times (j-1)!}$$

so that $e^{-m}\frac{m^6}{6!} = e^{-m}\frac{m^5}{5!}\frac{m}{6} = P_5\frac{m}{6} = 0\cdot 1280\frac{m}{6} = 0\cdot 0731$, and so on.

After this, the next column, column 6, showing values of E is worked out by multiplying the total number of frequencies, 63, by the probabilities in turn. In the 'In' table this was done directly by taking p_0 as $63e^{-m}$. In the last column of each of the two parts of the table, χ^2 is worked out from eqn. 11. Several of the lines have been grouped together to comply with the requirement that the expected frequency should be 6 or more. In connection with the significance of χ^2 the table will have two fewer degrees of freedom than the number of groups used for the calculation of χ^2, after the grouping mentioned in the last sentence has been done, so that here both tables will have four degrees of

freedom. One degree of freedom has been lost in calculating the mean m, and another is lost because when one less than the number of groups has been filled in, the last one follows from the general total. Here we see that the 'Out' table has only a small value of χ^2 giving a probability of between 0·8 and 0·7, and so it is clearly not significant. There is thus no reason to think that the traffic counted in this direction was inconsistent with the Poisson distribution.

On the other hand, the 'In' table has a large value of χ^2 with a probability of less than 0·01, which is highly significant. So that in this direction the flow of the traffic was not consistent with the Poisson distribution. On examination of the f column, column 8 of Table 24, we see that there was a deficiency in the number of times three vehicles passed in the interval, an excess in the number of times either none or one vehicle passed, and also in the higher numbers.

A probable explanation of this is that the stretch observed is a partly built-up and slightly winding level stretch of road at an altitude of about 400 feet above sea level, approached at both ends by gradients rising from about sea level. The observer whose figures were used was posted at the eastern end of the stretch, and vehicles passing in the 'In' direction, going westwards, had just come up one of these gradients, which has a long length in the single figures, about 1 in 7 (14%). Thus there would be a tendency for the faster vehicles to be unable to pass slower ones, with some bunching in consequence. In the other direction, the cars would have passed up a much longer and flatter gradient, and also along the more level stretch, so the traffic could sort itself out more easily, and so conform better to the random distribution.

The last example merely shows how traffic does, or does not, follow the Poisson distribution. The next one shows how this information can be put to practical use. It arose because a scheme was being prepared for improving a complicated junction in a small town. There was not room for a full-sized roundabout without very extensive, and expensive, demolition of buildings, but it was possible to provide a small one. From a preliminary study of the scheme, it looked very much as if it would work if entry from the side roads could be controlled by traffic lights. There was however, one feature of the layout which might have introduced the possibility of the roundabout jamming frequently. That would obviously have been a fatal objection to the scheme. The layout is shown diagrammatically in Fig. 13, on which the entering roads are lettered from A to F.

Roads B and C are one-way streets in the directions shown by the arrows. There is no difficulty in providing adequate weaving distances for the roads from B round to E, but there is difficulty if roads A and F

both enter the roundabout, because they converge very sharply. It is possible to run road F into road A as shown in the figure, and then to control the junction by traffic lights working in conjunction with those controlling the other roads. But the distance d, between the junction of F and the roundabout, cannot be arranged to allow more than about

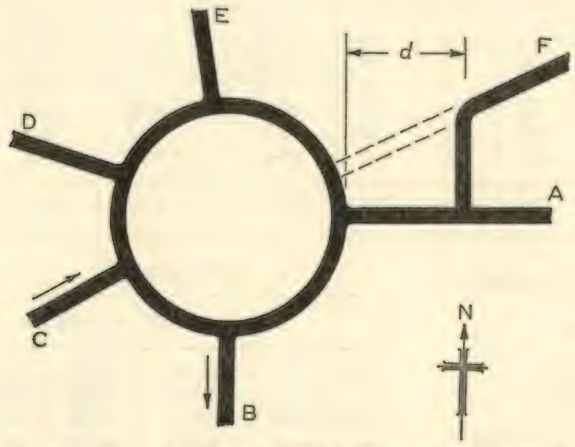


Fig. 13. Diagram of proposed roundabout with its entering roads

five or six west-bound vehicles to stand in it when the lights prevent their crossing the mouth of road F. More than that number would mean that some would have to wait on the roundabout itself, and so probably jam it. So to find out if the scheme is at all feasible, it is necessary to know the probability of more than five or six cars entering road A in the period when the lights would be set against them to enable traffic from road F to enter road A.

To find this out, a count was taken of the traffic which entered roads A and F from roads C and E, which the scheme envisages as being allowed to enter the roundabout together. The method was the same as in the last example. The recording of the registration numbers of the vehicles in this case enabled their movements to be traced. The result is shown in Table 25. In this table, column 1 shows the number of vehicles passing in the interval, which was half a minute. Column 2 shows the observed numbers passing in the interval, and the other columns have the same meanings as before. The working is shown to slide rule accuracy this time. In the last example the working was taken to more places than was really necessary, the better to illustrate the working.

The table shows that the agreement with the Poisson distribution is

TABLE 25. THE POISSON SERIES: TRAFFIC COUNT AT A JUNCTION

1	2	3	4	5	6	7	8
No. of vehicles	f (O)	fx	fx^2	E	$\frac{(O-E)^2}{E}$	$m=1{\cdot}5$ E	$m=2{\cdot}0$ E
7							0·3
6						0·1	1·4
5				0·3		1·6	4·1
4	4	16	64	1·5 } 8·1	0·15	5·4	10·3
3	3	9	27	6·3		14·2	20·7
2	22	44	88	20·0	0·20	28·8	31·1
1	40	40	40	42·3	0·12	38·4	31·1
0	46	0	0	44·6	0·04	25·6	15·6
$S(\ )$	115	109	219	115·0	0·51	115·1	114·6

$$\bar{x} = \frac{109}{115} = 0{\cdot}948$$

$$fx^2 = 219$$

$$\bar{x}S(x) = \frac{103{\cdot}3}{115{\cdot}7}$$

$$s^2 = \frac{115{\cdot}7}{114} = 1{\cdot}02$$

$$m = \bar{x} = 0{\cdot}948$$

$$e^{-m} = 0{\cdot}388$$

$$N = 2$$

$$\chi^2 = 0{\cdot}51$$

$p = 0{\cdot}8$: not significant

Probabilities

	$m=0{\cdot}948$	$m=1{\cdot}5$	$m=2{\cdot}0$
$p_0 = e^{-m}$	0·388	0·223	0·135
$p_1 = p_0 m$	0·368	0·334	0·270
$p_2 = p_1 m/2$	0·174	0·250	0·270
$p_3 = p_2 m/3$	0·055	0·125	0·180
$p_4 = p_3 m/4$	0·013	0·047	0·090
$p_5 = p_4 m/4$	0·003	0·014	0·036
$p_6 = p_5 m/6$	—	0·004	0·012
$p_7 = p_6 m/7$	—	—	0·003
	1·001	0·997	0·995

close, and that the probability p_5 of five vehicles entering is 0·3 times in the 115 intervals, about one hour, which was the busiest hour of the day. This is a risk which could be taken, but there is the possibility that the traffic might increase. As counted, the mean was 0·95, and so if the traffic increased by 50%, to 172 vehicles per hour, the mean would become approximately 1·5, and if it doubled the mean would become 2·0. The expected numbers were then worked out using these two values for the mean in columns 7 and 8 of the table, the probabilities being done in the panel at the bottom right of the table. Column 7 shows that the probable number of times five vehicles might wait in the half-minute interval is about 1·6 times in one hour if the traffic increases by 50%, and column 8 shows that it would be about four times in one hour if it doubled. As it so happens, the junction is on a temporary diversion of through traffic, pending the construction of a by-pass for the town, and the reasonable expectation is that this would be provided before the

traffic doubles, in which case the traffic passing through the junction would not increase so much. There was, therefore, a fair prospect that the scheme would work at first, and possibly even if the traffic increased by 50%, although it was doubtful if it would do so if the traffic doubled. Nevertheless, the analysis showed that the idea would be worth looking into further. If the layout could be altered so that six vehicles could wait, or if some form of filter could be provided into road F, the system could reasonably be expected to work for a number of years; so it was decided to investigate the matter in greater detail.

CHAPTER 14

The Exponential Distribution

We have seen in the last chapter that the number of events occurring in consecutive equal periods of time will have the Poisson distribution if they occur randomly in time. But the distribution of the intervals between the events, on the other hand, will not follow the Poisson distribution, but the *exponential distribution.* This is less frequently dealt with in textbooks, but it has its uses in traffic studies for such things as studies of waiting times of pedestrians crossing a road, or vehicles waiting at traffic lights. An outline of its derivation will be given here, because this is less familiar.

Let us suppose that there is a probability, denoted by p, of an event occurring in a short period of time, which is denoted by Δ. Then the probability p_j of there being j of these short periods between two consecutive events is the probability of j consecutive periods with no events occurring, followed by one period in which it does occur. That is

$$p_j = (1-p)^j p$$

If the events occur randomly in time, and can occur at any time, then p is approximately proportional to the length of the short period Δ, so if we put $\lambda = p/\Delta$, and if Δ is small enough for the probability p^2 of two events in the same period to be negligible, then λ should be approximately the same for any short period Δ. The probability of a time interval t between two consecutive events is approximately the probability that there are t/Δ consecutive 'no event' periods between, that is,

$$p_{(t)} = p_{t/\Delta} = (1-\lambda\Delta)^{t/\Delta} . \lambda\Delta$$

Strictly speaking, $p_{(t)}$ is the probability of a time interval within a small range on either side of t. Then letting t approach zero

$$\frac{p_{(t)}}{\Delta} \to \lambda e^{-\lambda t}$$

The limit of this probability $p_{(t)}$ divided by the width of the range is a probability density. Denoting this by $f_{(t)}$ we have that

$$f_{(t)} = \lambda e^{-\lambda t} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (45)$$

This expression is the frequency function of intervals between consecutive random events, and it defines the exponential distribution. In it, λ will be some function such as the rate of flow of traffic, because when vehicles can overtake freely they will be approximately randomly spaced.

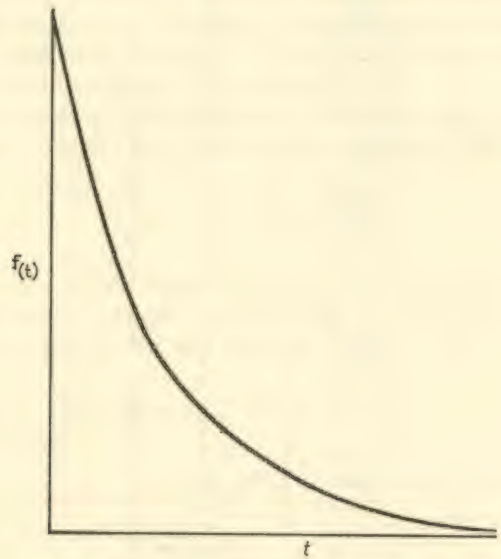


Fig. 14. The exponential distribution

The distribution is shown graphically in Fig. 14. Thus the equation gives a continuous distribution, like the normal. It does not violate the condition that the sum of the probabilities is unity. This sum is clearly the integral of the right-hand side of eqn. 45 within the limits of zero to positive infinity, that is, it is the area of the curve in Fig. 14. So

$$\int_0^\infty \lambda e^{-\lambda t}\, dt = \lambda\left(\frac{1}{\lambda}\right) = 1$$

In practice, we do not need to evaluate eqn. 45 directly. The reason for this will be illustrated by the next example, which is taken from the figures for the traffic count used in the last chapter, in Table 24 (page 103), for the 'Out' traffic. In this count the times of passing the observer were recorded within 0·01 of a minute, so the intervals between the pass-

ings could be obtained by subtraction. The frequency of intervals falling into equal groups of 0·10 of a minute is given in Table 26.

TABLE 26. THE EXPONENTIAL DISTRIBUTION: INTERVALS OF TIME BETWEEN VEHICLES PASSING

Interval, min	No. observed f
0·0–0·1	54
0·1–0·2	23
0·2–0·3	16
0·3–0·4	10
0·4–0·5	16
0·5–0·6	16
0·6–0·7	12
0·7–0·8	8
0·8–0·9	8
0·9–1·0	7
1·0–1·1	4
1·1–1·2	5
1·2–1·3	4
1·3–1·4	5
1·4–1·5	1
1·5–1·6	2
1·6–1·7	0
1·7–1·8	3
1·8–1·9	1
1·9–2·0	2
2·0–2·1	2
2·1–2·2	2
2·2–2·3	0
2·3–2·4	0
2·4–2·5	1
2·5–2·6	0
2·6–2·7	0
2·7–2·8	1
2·8–2·9	1
2·9–3·0	0
.....	.....
3·7–3·8	1
$S(\ )$	205

There are only 205 intervals in this table, as against 216 vehicles recorded in Table 24, because the observer missed the exact passing times of a few vehicles, and also did not record the times of passing of one or two buses which stopped in the stretch under observation. It has already been mentioned that the object of the timing was the average speed of vehicles over the stretch, although the count was in a form which could

be used in the analyses made in the last chapter, and in the present one. The buses would therefore not be of interest.

A glance at this table shows that, as would be expected from Fig. 14, most of the intervals are on the short side, and that there is a long tail of longer intervals. In this tail, also, many of the groups contain no observations. The whole table is very cumbersome. Yet if a wider grouping was adopted, the distribution would be lost altogether. Even with the small unit of grouping adopted, 0·1 minute, a quarter of the observations are in the first group. If double the unit of grouping was used, one-fifth of a minute, over one-third of the observations would be in the first group, and half of them in the first two, yet we would still be left with a cumbersome table of 17 groups.

We can get over this, and save work in testing the significance of the distribution, by using group boundaries calculated from the equation

$$t_i = -\frac{1}{\lambda}\log_e\left(\frac{i}{j}\right) \quad . \quad . \quad . \quad . \quad . \quad . \quad (46)$$

in which t_i is a group boundary, i = successively 0, 1, 2, 3, . . . j, and j is the number of groups required. This is because the expected frequency in any group, say that between t_i and $t_{(i+1)}$, is

$$\int_{t_{(i+1)}}^{t_i} \lambda e^{-\lambda t}\,dt = e^{-\lambda t_i} - e^{-\lambda t_{(i+1)}}$$

$$= \exp\left[\log_e\left(\frac{i+1}{j}\right)\right] - \exp\left[\log_e\left(\frac{i}{j}\right)\right]$$

$$= \frac{1}{j}$$

From this it will be seen that, if these boundaries are used, the expected frequency in all the groups will be $1/j$th of the total number in the sample, that is, it will be n/j, which greatly simplifies the arithmetic. Thus, if we use ten groups in the figures of Table 25, the expected frequency in each group will be $205/10 = 20{\cdot}5$.

We may now work out the group intervals to be used in our example. The exact mean period is not known directly in this case, owing to some vehicles having been missed, as mentioned above, but we can work out the mean interval from Table 26. This is 0·6078 min, and so the total time is $205 \times 0{\cdot}6078 = 124{\cdot}6$ min. Thus λ, which is here the rate of passing, becomes $206/12460 = 0{\cdot}01653$ vehicles per hundredth of a minute and $1/\lambda = 60{\cdot}5$.

Choosing ten intervals as a convenient number, we then multiply 60·5 by the values of $-\log_e(i/10)$ for the successive values of i from 1 to 10.

For convenience, values of $-\log(i/j)$ have been given in Table B at the end of the book for five convenient numbers of groups. So in this case multiplying 60·5 by the values of the log given in that table for $j = 10$ gives the boundaries as 0, 0·064, 0·135, 0·216, 0·309. 0·419, 0·554, 0·728, 0·974, 1·393 and infinity. The boundaries are taken to one more place than in the figures used, so that it will be clear which groups the figures enter.

The intervals are then counted afresh from the data with these group boundaries, and the result is set out in Table 27. In the ordinary way,

TABLE 27. THE EXPONENTIAL DISTRIBUTION: INTERVALS OF TIME BETWEEN VEHICLES PASSING (LOGARITHMIC GROUPING)

Group	O	$(O-E)^2$
0·00–0·06	35	210·2
0·07–0·13	28	56·3
0·14–0·21	22	2·2
0·22–0·30	11	90·3
0·31–0·41	11	90·3
0·42–0·55	25	20·2
0·56–0·72	21	0·3
0·73–0·97	18	6·2
0·98–1·39	16	20·3
>1·39	18	6·2
$S(\ )$	205	502·5

$$E = \frac{205}{10} = 20{\cdot}5$$

$$\chi^2 = \frac{S(O-E)^2}{E}$$

$$= \frac{502{\cdot}5}{20{\cdot}5} = 24{\cdot}5$$

for $N = 8$, $p < 0{\cdot}01$; highly significant

of course, only one count from the data would be necessary. The extra one has been put in here as illustration of the method, although in this case it had to be used to work out the mean interval. The data had not been gathered with this test in mind.

In every row, the expected frequency E is 20·5, as already found. Then O is the observed frequency, as counted from the data, so that $(O-E)^2$ can be calculated directly in each line, and the sum of the column divided by 20·5 to give χ^2. This will have eight degrees of freedom, because we use one degree in making totals agree, and another in fitting the distribution. In this case p is very low, and so we cannot conclude that the intervals here followed the exponential distribution. The table shows that this is accounted for by a much larger number of the small intervals than would be expected, and also deficiences in the 0·22/0·30 and 0·31/0·41 groups. There was evidently more bunching of vehicles than would be expected.

There appears to be an inconsistency between the results we have obtained in these figures, because in the previous chapter we found that

their agreement with Poisson was good, yet we now find that the fit with the exponential is poor. We have, however, been testing two different aspects of the figures. There could be a fair amount of bunching without affecting the distribution of counts in consecutive intervals of time, which was the effect tested in the last chapter. The present analysis would be more sensitive to it. The Poisson test is a weaker one, and really only uses part of the information, though this may still be of practical value.

CHAPTER 15

Sampling

Sometimes the figures for analysis will be ready to hand, as in those of Tables 15 and 18 (pages 71 and 83), in which the time intervals depend on the way that the original figures were set out. In many other cases we have some choice in the material to be used. In the speed timings we can choose what vehicles to time, where to time them, and how many to time. When we have this amount of control, it is essential that the sample we choose should be *random:* that is, each individual possible observation in the data we are examining should have an equal chance of being included in the sample. If there is any tendency for some effect or other to make it more likely that some observations would be chosen in preference to others, this tendency would produce a *bias.*

It is not always easy to set up a truly random sample, and bias may very easily creep in. An example of interest to traffic engineers, often quoted, is a traffic survey in the U.S.A., in which householders were to be questioned as to their motoring habits. To do this, it was decided to visit every tenth house, taken from a street list, normally a fairly unexceptionable way of doing this kind of thing. It was then noticed in the analysis that there seemed to be an abnormal number of shopkeepers, and it was realized that the gridiron layout of the city, which is common in the U.S.A., meant that in this particular case every tenth house tended to be a corner plot, and so a shop, which are often on corner sites. Another case is cited by Tippett[5] of a study to find out the effect of giving milk to schoolchildren. The actual children to be given milk were individually chosen by a correct method, but then the teachers were given some discretion to vary the actual children to be given the milk. This introduced a bias, because the teachers, somewhat humanly, tended to favour the weaker-looking children, and give them the milk, in preference to the stronger child who had been chosen in the sampling process.

Another classic case, also mentioned by Tippett, is the straw vote which was carried out to forecast the result of the 1936 Presidential election in the U.S.A. This vote forecast that Roosevelt would be in a minority, whereas in fact he carried the election by a substantial majority. The vote was taken by sending cards to a sample of people chosen from the telephone directory, and from lists of car owners. In the circumstances of 1936, these people included a majority which would be likely to vote Republican, and so again a bias crept in.

Taking examples of methods used in practice, in interrogation traffic surveys, drivers are questioned about their intended directions, and the usual practice is for a policeman to direct into the survey point the next vehicle passing when the observers have signalled to him that they are ready. The danger to be guarded against here is that the policeman may tend to stop vehicles which head strings of traffic and let the faster ones go on, thus overweighting the survey with commercial vehicles.

In Dorset, when origin and destination surveys are carried out, the registration number method is used. In this, the registration numbers of vehicles passing are recorded, either by writing them each on a plain card, or—a more recent experiment—they are 'mark-sensed' straight on to a special card for electrical punched card analysis. In the former method the cards have to be compared manually, which is laborious and therefore costly. But even in the punched card system, although the cards for analysis are cut direct from the field card by machine, the cost of analysing a large number of cards is still high. So in either case a sampling technique is advisable.

It is also important that the car numbers chosen for the sample should be easy to read in the field, and not easily confused with other numbers. For instance, with the somewhat poor type of lettering used for number plates in this country, it is very easy to confuse 5 and 6. After some experimenting, the following system has been evolved, and found easiest to use. First, all buses and coaches are recorded. Then for private cars and lorries, with the three-letter and three-number type of plate—such as ABC 123—those with numbers from 1 to 199, and from 700 to 799, all inclusive, are recorded. It makes no difference whether the letters are first or last, but all combinations of letters associated with the chosen numbers are recorded. For the now relatively rare four-number and two-letter plates, the first number is treated as a letter, and the last three numbers then treated as a three number combination. Thus a car with the number AB 1234—or 1234 AB—would be treated as if its number was AB1 234, and not recorded. On the other hand, if its number had been AB 1734—or 1734 AB—it would be treated as AB1 734, and recorded.

This gives about a 30% sample in practice, and has been found to work well. One minor possibility of bias is that 'V.I.P.s' are apt to obtain short numbers like ABC 1, and if they are local might tend to be about more frequently. Consultation with the local licensing department has shown that this is probably not frequent enough to matter, and it does not seem to have happened in practice.

One of the commonest general methods of choosing random samples is to write the particulars of each observation on a separate card. The cards are then thoroughly mixed, and enough are dealt out to make a sample. Another way is to number all the cards, or observations, consecutively, and then choose the number needed with the help of a table of random numbers. Such a table is given by Lindley and Miller.[7]

In choosing a sample, we may need to know the size of sample we must use to get the accuracy we want. Having chosen and analysed it, we may want to know what degree of confidence we can have that it truly represents the population. These questions can be answered by using the principle, already stated (page 40) that the means of a number of random samples each of n observations, taken from a population which is normally distributed with variance σ^2, will themselves be normally distributed with variance σ^2/n. For convenience, the second of the two questions in the first part of this paragraph will be considered first.

The symbol σ_m will be used for the standard deviation of the means, and s_m for the root mean square of the sample means. As before, s will be used for the root mean square, the estimated standard deviation, of the population. From the principle mentioned above we first have that

$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

But we have to use the root mean square because we do not normally know the variance of the population, and so we have our equation for general use

$$s_m = \frac{s}{\sqrt{n}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (47)$$

As an example of the use of this expression, we may consider the data of Table 6 (page 35). In this table, we found that the root mean square s of the data was 7·004 miles/h, the mean was 32·3 miles/h and the number in the sample was 1660. Substitution in eqn. 47 gives

$$s_m = \frac{7{\cdot}004}{\sqrt{1660}} = \frac{7{\cdot}004}{40{\cdot}73} = 0{\cdot}172 \text{ miles/h}$$

It has been mentioned previously that about two-thirds of the observations in the normal distribution lie within the area of the curve bounded by the mean $\pm$ one standard deviation σ in Fig. 5. The more accurate figure is 68·3%. The means of the sample, then, will lie within the limits given by the calculation above for about two-thirds—more precisely 68·3%—of the samples taken. We may thus have confidence that for two-thirds of the samples the mean will lie within the limits $32{\cdot}3 \pm 0{\cdot}17$ miles/h, i.e. within the limits of about 32·5 and 32·1 miles/h. If this is wanted as a percentage error, it would be 0·17/32·3 or 0·53%, again with the two-thirds confidence.

This is a large sample, and so it may be of interest to see what the range would be for a smaller one, which will be taken as that in column 7 of Table 11 (page 57). This is one of 76 observations, with a mean of 25·4 miles/h, and a root mean square of $\sqrt{1{\cdot}015} = 1{\cdot}008 = 4{\cdot}03$ miles/h. Then

$$s_m = \frac{4{\cdot}03}{\sqrt{76}} = 0{\cdot}47$$

The range for the smaller sample is thus rather wider than that for the larger one and in this case we can have two-thirds confidence that the range lies between about 24·9 and 25·9 miles/h.

We may also want to know what is the probability of our sample coming from a population with some other mean, denoted by μ. To find this out, we can use the t test, and

$$t = \left| \frac{\bar{x} - \mu}{s_m} \right| \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (48)$$

Applying this to the last example, we may inquire what is the probability of the sample coming from a population with a true mean μ of 27 miles/h. Substituting in eqn. 48, we have

$$t = \frac{27{\cdot}0 - 25{\cdot}4}{0{\cdot}47} = \frac{1{\cdot}6}{0{\cdot}47} = 3{\cdot}40$$

For 75 degrees of freedom, we use the table of normal deviates, and we find that p is less than 0·001. It is therefore highly improbable that the sample will come from a population with a mean of 27 miles/h.

We may again make use of the t test to find the *confidence limits* for the mean of the population. These are the two values within which we can expect the population mean to fall, within some selected degree of confidence. They are denoted by the symbol μ_p. From eqn. 48 it can easily be seen that the two values derived from this expression are

$$\mu_p = \bar{x} \pm t s_m \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (49)$$

which will give us our confidence limits. The degree of confidence will be that for which we have chosen t, allowing for the size of the sample.

Applying this to the same example as before, and using the 0·05 confidence level, for which t is very nearly 2 for a fairly large sample, say over 30, we have

$$\mu_p = 25{\cdot}4 \pm (2 \times 0{\cdot}47) = 25{\cdot}4 \pm 0{\cdot}94$$

So our two limits are 24·5 and 26·3, and we can be about 95% confident that the true mean of the population is likely to lie within these limits.

One important question is the size of the sample we must use. The cost of obtaining the sample may rise steeply with its size. If the traffic is light, for example, the observers may have to wait a long time to obtain a large number of observations. It would then be useful to know what size of sample would be needed to get some desired degree of accuracy. Let us assume that we do not want our result to differ from the true mean by more than a limit denoted by ε, again with some chosen degree of confidence. We can make $\varepsilon = \bar{x} - \mu_p$, and so the equation becomes

$$t = \frac{\varepsilon}{s_m} = \frac{\varepsilon}{\frac{s}{\sqrt{n}}} = \frac{\varepsilon\sqrt{n}}{s}$$

In this, n will be the number of observations in the sample, which for clearness would be better denoted by n_s. Then, substituting this symbol, squaring and rearranging, we get,

$$n_s = \left[\frac{ts}{\varepsilon}\right]^2 \quad \ldots\ldots\ldots \quad (50)$$

The use of this expression depends on having knowledge of the variance of the population. We can sometimes estimate this from other information available. For example, we find that for the timings within the speed limits in Table 11 (page 57) the combined mean square s^2 was 2·301 in working units. The working unit was 4 miles/h, and so we have to multiply by 16, which gives 36·8 in (miles/h)^2. To find the number of vehicles to time for a further test we might use a figure of 37 (miles/h)^2 for the variance to use in eqn. 50. Suppose further that we want to use the 0·05 confidence limit, and want a result that we can be confident is not more than 1 miles/h from the true mean. We would then find that

$$n_s = \frac{4 \times 37}{1} = 148, \text{ say } 150$$

We might, however, decide that the site conditions were more like those where the timings were made which were recorded in column 8 of the same table. In this case $s^2 = 1{\cdot}015$ in working units $= 16{\cdot}2$ (miles/h)^2, and then

$$n_s = \frac{4 \times 16}{1} = 64$$

We see that the estimate of the variance makes a quite considerable difference to the number of observations needed, but the equation does give some useful guide to the size of sample needed.

So far, the samples we have considered have dealt with the actual numbers, or frequencies, of the observations. But we may wish to express the result in terms of proportions or percentages. Thus, for example, we may want to know the proportion of the population—in the conventional sense of the word—of an area which uses public transport, or the proportion of traffic which will use one or other of two alternative routes. We then again have the problem of knowing what size of sample to use to get certain limits of confidence.

In this case, the notation of the binomial distribution (page 18) will be used, and p will denote one of the proportions—e.g. that using public transport, or that following one of the routes being investigated—which will, as before, be called a success. The other proportion will be denoted by q, and then $(p+q) = 1$.

Then the probability that our first observation will be a success is also p, and if we denote the total population by P, the total number of successes will be pP. When a second observation is made, the probability of its being a success will be either $(pP-1)/(P-1)$ or $pP/(P-1)$ according to the result of the first observation. But clearly if the population is large the probability is very nearly p in either case. So provided that the sample size n is small in proportion to the size of the population P, very little error will result from two assumptions: first that the probability that any one observation will be a success is p, and second that this is very nearly independent of the results of the observations already made.

Thus, if we have a sample of size n, the probability p_r that there will be r successes in the sample of n observations is given by the general term of the binomial expansion,

$$p_r = \frac{n!}{r!(n-r)!} p^r q^{(n-r)}$$

The binomial distribution has a mean of np, and variance npq, and provided p is not too small on the one hand, or too near to unity on the other, then, for moderately large samples, the normal distribution is a

fairly good approximation to the binomial. So if p is known, we say that there is about a 95% probability that the observed r is within two standard deviations of the real mean, so that

$$|r-np| \leqq 2\sqrt{(npq)} \quad . \quad . \quad . \quad . \quad . \quad . \quad (51)$$

This is derived from the property of the normal deviate (page 41) that the probability of getting a normal deviate c of 2—or a deviation of twice the standard deviation—is very nearly one in twenty, or 5%. In eqn. 51, $|r-np|$ represents $|d-\mu|$ and $\sqrt{(npq)}$ is σ.

But the purpose of the survey is to estimate p as a proportion, and while we can take r/n as an estimate of it from the sample in a similar fashion to that in which we take m, the mean of the sample, as an estimate of the mean μ of the population, we want to know how accurately we can expect the estimate to be. We do not know, but by using eqn. 51 we can derive a range of values of p which we can have approximately 95% confidence will contain the true value of the proportion p. To find the limits of the range of p, we make two sides of eqn. 51 equal, and by squaring both sides, and dividing throughout by n^2, we get

$$\left(\frac{r}{n}-p\right)^2 = 4p\frac{q}{n}$$

whence

$$p^2\left(1+\frac{4}{n}\right)-p\frac{2}{n}(r-2)+\frac{r^2}{n}=0 \quad . \quad . \quad . \quad . \quad (52)$$

The two solutions of this quadratic equation are our upper and lower confidence limits for p.

A simpler method of finding out the approximate confidence limits is obtained by substituting r/n for p in the right-hand side of eqn. 51. Dividing through by n, we then get for our limits,

$$p = \frac{r}{n} \pm 2\sqrt{\left[\frac{\left(1-\frac{r}{n}\right)\frac{r}{n}}{n}\right]} \quad . \quad . \quad . \quad . \quad . \quad (53)$$

Unless either r or $(n-r)$ is very small, the use of eqn. 53 will give a satisfactory result.

At a particular Y-junction, a first-class road diverges from a trunk road, so that traffic travelling westwards has two alternative routes to take, while the traffic travelling eastwards must join together on to the trunk road. It was wanted to know in what proportion the westbound traffic is drawn off the trunk road to the west of the junction by the first-class road or the eastbound traffic is contributed to by it, as the

case may be. A two-hour count was taken, and it was found that the traffic on the undivided trunk road east of the junction was 1202 vehicles. Of these, 34·4% were on the trunk road west of the junction, and 65·6% on the first-class road; suppose we want to know what are the limits of confidence of the two proportions. Calling a vehicle on the trunk road west of the junction a success, we may then put $r/n = 0{\cdot}344$, and n is of course 1202. Substituting in eqn. 53, we have

$$p = 0{\cdot}344 \pm 2\sqrt{\left[\frac{0{\cdot}656 \times 0{\cdot}344}{1202}\right]}$$
$$= 0{\cdot}344 \pm 0{\cdot}027 = 34{\cdot}4\% \pm 2{\cdot}7\%$$

So we may conclude that we may have 95% confidence that the percentage of traffic on the trunk road west of the junction is within the range 37·2% and 31·6%.

Before conducting any survey, the size of sample necessary to give the required accuracy should be approximately determined. For a survey to estimate a proportion or percentage, the standard deviation of the ratio r/n is $\sqrt{(pq/n)}$; so to get an answer from the survey which we can be, say, 95% confident is within an amount ε of the true value, we need n to be large enough for $2\sqrt{(pq/n)}$ to be less than ε. Therefore

$$n \geqq 4pq/\varepsilon \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (54)$$

but this involves the use of p, which is unknown; so either an estimated value, or one obtained from a small pilot survey, should be used.

Table C, at the end of the book, gives approximate sample sizes necessary to estimate p to within an accuracy of 100ε%, when $100p$% is within certain ranges, for 95% confidence limits.

In the last example, it was suspected that the traffic was divided between about one-third to the trunk road, and two-thirds to the first-class road; so p would be expected to be within the 30%–70% range in the table, and for a 5% accuracy the table indicates that a sample of about 400 would be needed.

CHAPTER 16

Derivation of the Algebraic Expressions

1. *Derivation of* $S(x-\bar{x})=0$ (*see page* 24)

This identity follows from the definition of the mean, but its derivation algebraically illustrates the general method in the simplest way; thus it will be useful, to explain what follows.

Assume a sample of n observations, respectively x_1, x_2, x_3, x_4, and so on up to x_n. We then derive the value of $S(x-\bar{x})$ by adding up the individual values of $(x-\bar{x})$; thus

$$\begin{array}{l} x_1-\bar{x} \\ x_2-\bar{x} \\ x_3-\bar{x} \\ \ldots\ldots \\ \ldots\ldots \\ x_n-\bar{x} \\ \hline S(x)-n\bar{x}=S(x)-S(x)=0 \end{array}$$

Adding

because, by definition, $x_1+x_2+x_3+\ldots+x_n=S(x)$, and from eqn. 4 $n\bar{x}=S(x)$.

2. *Derivation of the Expression for* $S(x-\bar{x})^2$ (*see page* 30)

Again assume a sample of $x_1, x_2, x_3 \ldots x_n$. Then proceed as before, but first square each of the deviations from the mean, and again add up the results,

for

$$\begin{array}{lll} x_1 & \ldots\ldots\ldots & (x_1-\bar{x})^2=x_1^2-2x_1\bar{x}+\bar{x}^2 \\ x_2 & \ldots\ldots\ldots & (x_2-\bar{x})^2=x_2^2-2x_2\bar{x}+\bar{x}^2 \\ x_3 & \ldots\ldots\ldots & (x_3-\bar{x})^2=x_3^2-2x_3\bar{x}+\bar{x}^2 \\ \multicolumn{3}{c}{\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots} \\ \multicolumn{3}{c}{\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots} \\ x_n & \ldots\ldots\ldots & (x_n-\bar{x})^2=x_n^2-2x_n\bar{x}+\bar{x}^2 \end{array}$$

Adding

$$S(x-\bar{x})^2=S(x^2)-2\bar{x}S(x)+n\bar{x}^2$$

because both $\bar{x}$ and $\bar{x}^2$ are the same for any value of x, and so are constants in the summation. But, from eqn. 4 the algebraic sum of the two last terms can become either $-S^2(x)/n$ or $-\bar{x}S(x)$ at choice, giving the two forms of eqn. 6.

3. *Derivation of the Summation Method for* $S(fx^2)$ (*see page* 34)

Again assume a sample of n observations, as in the last two sections, and let the group frequencies be a, b, c, d, etc., as shown in the table below, which represents a grouped table algebraically, with x taken from zero in the lowest group to j in the top group.

1	2	3	4	5	6
x	f	Standard method		Summation method	
		fx	fx^2	C1	C2
j	a	aj	aj^2	a	a
$j-1$	b	$b(j-1)$	$b(j-1)^2$	$a+b$	$2a+b$
$j-2$	c	$c(j-2)$	$c(j-2)^2$	$a+b+c$	$3a+2b+c$
$j-3$	d	$d(j-3)$	$d(j-3)^2$	$a+b+c+d$	$4a+3b+2c+d$
....	....	.....	.....	.....	.....
....	....	.....	.....	.....	.....
0		0	0	0	0
	$S(f)$	$S(fx)$	$S(fx^2)$	$aj+b(j-1)$ $+c(j-2)$ $+d(j-3)$ etc. $=$ S(C1)	$a\{1+2+3+4+\ldots+j\}$ $+b\{1+2+3+4+\ldots+(j-1)\}$ $+c\{1+2+3+4+\ldots+(j-2)\}$ etc. $=$ S(C2)

Columns 2 to 4 show the standard method in which the frequency is multiplied by its x value j, $(j-1)$, $(j-2)$ etc., to get fx and fx^2. Columns 5 and 6 show the corresponding summation method. In column 5, the C1 column, the frequencies are summed cumulatively from the top downwards, giving successive terms of a, $(a+b)$, and so on. Comparison of this column with the fx column shows that its sum is $S(fx)$. Then, when this C1 column is again summed cumulatively from the top in the C2 column, column 6, we get a in the first line, and $(a+a+b)=(2a+b)$ in the next column, and so on. Then when this column is added up, we get a $(1+2+3+\ldots+j)$ in the first term and each successive term in b, c, etc., is the same series, but each ending in $(j-1)$, $(j-2)$, etc., successively down the column.

Now each of these coefficients of the frequencies a, b, etc., is a series of the form $(1+2+3+\ldots+j)$. The sum of this series is $j(j+1)/2$. If we

double this we get $j(j+1)=j^2+j$. Then by subtracting j from this we get j^2. By extension therefore, if we sum the C2 column, double it, and subtract the sum of the C1 column, we get $S(fx^2)$.

4. *Analysis of variance. Derivation of Table* 10 (*page* 54)

Referring to Fig. 8 (page 53), we see that

$$(x-\bar{X})=(x-\bar{x}_a)+(\bar{x}_a-\bar{X})$$

Squaring

$$(x-\bar{X})^2=(x-\bar{x}_a)^2+2(x-\bar{x}_a)(\bar{x}_a-\bar{X})+(x_a-\bar{X})^2$$

As before, we may consider this to be representative of any term of a sample composed of terms x_1, x_2, etc., to x_n, and these terms may be thought of as forming an array of n terms. If we then sum all these terms in an array, we have

$$S_a(x-\bar{X})^2=S_a(x-\bar{x}_a)^2+2(\bar{x}_a-\bar{X})\,.\,S(x-\bar{x}_a)+n_a(\bar{x}_a-\bar{X})^2$$

For any one array, $(\bar{x}_a-\bar{X})$ is a single measurement and is the distance of the array mean from the grand mean. $S(x-\bar{x}_a)$ is, as usual, zero, so that the expression becomes

$$S_a(x-\bar{X})^2=S(x-\bar{x}_a)^2+n_a(x_a-\bar{X})$$

Then for the grand array, by summing all the arrays, we get

$$S(x-\bar{X})^2=SS_a(x-\bar{x}_a)^2+S[n_a(x_a-\bar{X})^2]$$

This expression is represented in the third column of Table 10.

5. *Regression. Derivation of eqn.* 17 *and* 18 (*page* 66)

The problem is to minimize the sum of squares of the deviations of the observations of y from the corresponding values derived from the regression equation, i.e. $S(y-Y)^2$.

To do this, we first of all subtract both sides of eqn. 16 from y. This gives

$$y-Y=y-a-b(x-\bar{x})$$

Squaring

$$(y-Y)^2=y^2+a^2-2ay+b^2(x-\bar{x})^2-2by(x-\bar{x})+2ab(x-\bar{x})$$

Summing this up for the whole of the observations in the usual way, we have

$$S(y-Y)^2$$
$$=S(y^2)+na^2-2aS(y)+b^2S(x-\bar{x})^2-2bS[y(x-\bar{x})]+2abS(x-\bar{x})$$

The last term is zero.

To find the minimum values of a and b, this expression is then partially differentiated with respect to a and b in turn, and the results equated to zero. First, partially differentiating with respect to a, and equating to zero, we have

$$\frac{\partial}{\partial a}[S(y-Y)^2]=2an-2S(y)=0$$

So

$$a=\frac{S(y)}{n}=\bar{y}$$

This is eqn. 17.

Next, partially differentiating with respect to b, and equating to zero, we have

$$\frac{\partial}{\partial b}[S(y-Y)^2]=2bS(x-\bar{x})^2-2S[y(x-\bar{x})]=0$$

So

$$b=\frac{S[y(x-\bar{x})]}{S(x-\bar{x})^2}$$

This is eqn. 18.

6. *The Deviation Cross-Product* (*for calculating the covariance*) (*see page* 66)

Any individual term in one form of the deviation cross-product is $(x-\bar{x})(y-\bar{y})$. Multiplying this out gives

$$(x-\bar{x})(y-\bar{y})=x(y-\bar{y})-\bar{x}y+\bar{x}\bar{y}$$

Summing this for the n terms gives

$$S[(x-\bar{x})(y-\bar{y})]=S[x(y-\bar{y})]-\bar{x}S(y)+n\bar{x}\bar{y}$$

The last two terms cancel, so that

$$S[(x-\bar{x})(y-\bar{y})]=S[y(x-\bar{x})]$$

A similar method can be used to derive the other form.

To derive eqn. 19 (page 66), the expression for calculating the deviation cross-product, again multiply out in a different form, and we get

$$y(x-\bar{x})=xy-y\bar{x}$$

Again summing for the n terms,

$$S[y(x-\bar{x})]=S(xy)-\bar{x}S(y)$$
$$=S(xy)-n\bar{x}\bar{y}$$

TABLES

Reference 7 in the Bibliography gives two useful books of tables, and Reference 6 also gives tables in some detail.

TABLE A. GROUP VALUES OF z IN THE EQUATION $z=\frac{1}{2}[\log_e P-\log_e(1-P)]$

(see page 22)

	z	Proportion P
+	2·00–1·75	0·98–0·99
	1·75–1·50	0·96–0·97
	1·50–1·25	0·93–0·95
	1·25–1·00	0·89–0·92
	1·00–1·75	0·83–0·88
	0·75–0·50	0·74–0·82
	0·50–0·25	0·63–0·73
	0·25–0·00	0·50–0·62
−	0·00–0·25	0·50–0·38
	0·25–0·50	0·37–0·27
	0·50–0·75	0·26–0·18
	0·75–1·00	0·17–0·12
	1·00–1·25	0·11–0·08
	1·25–1·50	0·07–0·05
	1·50–1·75	0·04–0·03
	1·75–2·00	0·02–0·01

Note: All the values of P in the table are inclusive, except 0·50. That is, values of P of 0·83 to 0·88 inclusive should be put into the +1·00 to +1·75 group of z, and so on. Observations of P = 0·50 should be divided equally between the 0·00 to +0·25 and the 0·00 to −0·25 groups of z.

TABLE B. VALUES OF $-\log_e(i/j)$ IN EQUATION 46 FOR GROUPING IN THE EXPONENTIAL DISTRIBUTION

(see pages 110*ff*.)

j \ i	12	11	10	9	8	7	6	5	4	3	2	1	0
8	—	—	—	—	0	0·1335	0·2877	0·4700	0·6931	0·9808	1·3863	2·0794	∞
9	—	—	—	0	0·1178	0·1823	0·4055	0·5877	0·8109	1·0986	1·5041	2·1972	∞
10	—	—	0	0·1054	0·2231	0·3567	0·5108	0·6931	0·9163	1·2040	1·6094	2·3026	∞
11	—	0	0·0953	0·2007	0·3185	0·4520	0·6061	0·7884	1·0116	1·2993	1·7047	2·3979	∞
12	0	0·0870	0·1823	0·2877	0·4055	0·5390	0·6931	0·8755	1·0986	1·3863	1·7918	2·4849	∞

TABLE C. APPROXIMATE SIZE OF SAMPLE NECESSARY FOR 95% CONFIDENCE THAT AN ESTIMATE OF A PROPORTION p IS ACCURATE TO WITHIN $100\varepsilon\,\%$

Range of $100p$, % / 100ε, %	30–70	20–30 70–80	15–20 80–85	10–15 85–90	7–10 90–93	5–7 93–95	4–5 95–96
½	40000	30000	25000	20000	15000	10000	7500
1	10000	7500	6000	5000	4000	2500	2000
2	2500	2000	1500	1200	1000	600	500
5	400	300	250	200	150	120	100

TABLE D. VALUES OF e^{-m}

The differences are subtracted. At underlined entries the interval in the value of m is changing.

m	e^{-m}	Diff.	m	e^{-m}	Diff.	m	e^{-m}	Diff.	m	e^{-m}	Diff.
0·00	1·0000		0·20	0·8187		0·40	0·6703		0·60	0·5488	
		100			81			66			55
0·01	0·9900		0·21	0·8106		0·41	0·6637		0·61	0·5433	
		98			81			66			54
0·02	0·9802		0·22	0·8025		0·42	0·6571		0·62	0·5379	
		98			80			66			53
0·03	0·9704		0·23	0·7945		0·43	0·6505		0·63	0·5326	
		96			79			65			53
0·04	0·9608		0·24	0·7866		0·44	0·6440		0·64	0·5273	
		96			78			64			52
0·05	0·9512		0·25	0·7788		0·45	0·6376		0·65	0·5221	
		94			77			63			52
0·06	0·9418		0·26	0·7711		0·46	0·6313		0·66	0·5169	
		94			77			63			52
0·07	0·9324		0·27	0·7634		0·47	0·6250		0·67	0·5117	
		93			76			62			51
0·08	0·9231		0·28	0·7558		0·48	0·6188		0·68	0.5066	
		92			75			62			50
0·09	0·9139		0·29	0·7483		0·49	0·6126		0·69	0·5016	
		91			75			61			50
0·10	0·9048		0·30	0·7408		0·50	0·6065		0·70	0·4966	
		90			74			60			50
0·11	0·8958		0·31	0·7334		0·51	0·6005		0·71	0·4916	
		89			73			60			49
0·12	0·8869		0·32	0·7261		0·52	0·5945		0·72	0·4867	
		88			72			59			48
0·13	0·8781		0·33	0·7189		0·53	0·5886		0·73	0·4819	
		87			71			59			48
0·14	0·8694		0·34	0·7118		0·54	0·5827		0·74	0·4771	
		87			71			58			47
0·15	0·8607		0·35	0·7047		0·55	0·5769		0·75	0·4724	
		86			70			57			47
0·16	0·8521		0·36	0·6977		0·56	0·5712		0·76	0·4677	
		84			70			57			47
0·17	0·8437		0·37	0·6907		0·57	0·5655		0·77	0·4630	
		84			68			56			46
0·18	0·8353		0·38	0·6839		0·58	0·5599		0·78	0·4584	
		83			68			56			46
0·19	0·8270		0·39	0·6771		0·59	0·5543		0·79	0·4538	
		83			68			55			45
0·20	0·8187		0·40	0·6703		0·60	0·5488		0·80	0·4493	

m	e^{-m}	Diff.	m	e^{-m}	Diff.	m	e^{-m}	Diff.	m	e^{-m}	Diff.
0·80	0·4493		1·00	0·3679		1·20	0·3012		1·60	0·2019	
		45			37			60			40
0·81	0·4448		1·01	0·3642		1·22	0·2952		1·62	0·1979	
		44			36			58			39
0·82	0·4404		1·02	0·3606		1·24	0·2894		1·64	0·1940	
		44			36			57			38
0·83	0·4360		1·03	0·3570		1·26	0·2837		1·66	0·1902	
		43			36			56			38
0·84	0·4317		1·04	0·3534		1·28	0·2781		1·68	0·1864	
		43			35			56			37
0·85	0·4274		1·05	0·3499		1·30	0·2725		1·70	0·1827	
		42			34			54			36
0·86	0·4232		1·06	0·3465		1·32	0·2671		1·72	0·1791	
		42			35			53			36
0·87	0·4190		1·07	0·3430		1·34	0·2618		1·74	0·1775	
		42			34			51			35
0·88	0·4148		1·08	0·3396		1·36	0·2567		1·76	0·1720	
		41			34			51			34
0·89	0·4107		1·09	0·3362		1·38	0·2516		1·78	0·1686	
		41			33			50			33
0·90	0·4066		1·10	0·3329		1·40	0·2466		1·80	0·1653	
		41			33			49			33
0·91	0·4025		1·11	0·3296		1·42	0·2417		1·82	0·1620	
		40			33			48			32
0·92	0·3985		1·12	0·3263		1·44	0·2369		1·84	0·1588	
		39			33			47			31
0·93	0·3946		1·13	0·3230		1·46	0·2322		1·86	0·1557	
		39			32			46			31
0·94	0·3907		1·14	0·3198		1·48	0·2276		1·88	0·1526	
		39			32			45			30
0·95	0·3868		1·15	0·3166		1·50	0·2231		1·90	0·1496	
		39			31			44			30
0·96	0·3829		1·16	0·3135		1·52	0·2187		1·92	0·1466	
		38			31			43			29
0·97	0·3791		1·17	0·3104		1·54	0·2144		1·94	0·1437	
		38			31			42			28
0·98	0·3753		1·18	0·3073		1·56	0·2102		1·96	0·1409	
		37			31			42			28
0·99	0·3716		1·19	0·3042		1·58	0·2060		1·98	0·1381	
		37			30			41			28
1·00	0·3679		1·20	0·3012		1·60	0·2019		2·00	0·1353	

m	e^{-m}	Diff.	m	e^{-m}	Diff.	m	e^{-m}	Diff.	m	e^{-m}	Diff.
2·00	0·1353		2·40	0·0907		3·40	0·0334		4·80	0·0082	
		27			44			16			8
2·02	0·1326		2·45	0·0863		3·45	0·0318		4·90	0·0074	
		26			42			16			7
2·04	0·1300		2·50	0·0821		3·50	0·0302		5·00	0·0067	
		25			40			15			6
2·06	0·1275		2·55	0·0781		3·55	0·0287		5·10	0·00610	
		25			38			14			58
2·08	0·1250		2·60	0·0743		3·60	0·0273		5·20	0·00552	
		25			36			13			53
2·10	0·1225		2·65	0·0707		3·65	0·0260		5·30	0·00499	
		25			35			13			47
2·12	0·1200		2·70	0·0672		3·70	0·0247		5·40	0·00452	
		24			33			12			43
2·14	0·1176		2·75	0·0639		3·75	0·0235		5·50	0·00409	
		23			31			11			39
2·16	0·1153		2·80	0·0608		3·80	0·0224		5·60	0·00370	
		23			30			11			35
2·18	0·1130		2·85	0·0578		3·85	0·0213		5·70	0·00335	
		22			28			11			32
2·20	0·1108		2·90	0·0550		3·90	0·0202		5·80	0·00303	
		22			27			10			29
2·22	0·1086		2·95	0·0523		3·95	0·0192		5·90	0·00274	
		21			25			9			26
2·24	0·1065		3·00	0·0498		4·00	0·0183		6·00	0·00248	
		21			24			17			
2·26	0·1044		3·05	0·0474		4·10	0·0166				
		21			23			16			
2·28	0·1023		3·10	0·0451		4·20	0·0150				
		20			22			14			
2·30	0·1003		3·15	0·0429		4·30	0·0136				
		20			21			13			
2·32	0·0983		3·20	0·0408		4·40	0·0123				
		20			20			12			
2·34	0·0963		3·25	0·0388		4·50	0·0111				
		19			19			11			
2·36	0·0944		3·30	0·0369		4·60	0·0100				
		19			18			9			
2·38	0·0925		3·35	0·0351		4·70	0·0091				
		18			17			9			
2·40	0·0907		3·40	0·0334		4·80	0·0082				

TABLE E. THE NORMAL DEVIATE, c

Probability	0·95	0·90	0·80	0·70	0·60	0·50	0·40
c	0·063	0·13	0·25	0·39	0·52	0·67	0·84
Probability	0·30	0·20	0·10	0·05	0·02	0·01	0·001
c	1·04	1·28	1·64	1·96	2·33	2·58	3·29

Table E is abridged from Table II of Fisher and Yates: 'Statistical Tables for Biological, Agricultural and Medical Research', published by Oliver and Boyd Ltd., Edinburgh, and reproduced by permission of the authors and publishers.

TABLE F. VALUES OF t

Degrees of freedom	t				
	0·10	0·05	0·02	0·01	0·001
1	6·31	12·71	31·82	63·66	636·62
2	2·92	4·30	6·97	9·93	31·60
3	2·35	3·18	4·54	5·84	12·94
4	2·13	2·78	3·75	4·60	8·61
5	2·02	2·57	3·37	4·03	6·86
6	1·94	2·45	3·14	3·71	5·96
7	1·90	2·37	3·00	3·50	5·41
8	1·86	2·31	2·90	3·36	5·04
9	1·83	2·26	2·82	3·25	4·78
10	1·81	2·23	2·76	3·17	4·59
11	1·80	2·20	2·72	3·11	4·44
12	1·78	2·18	2·68	3·06	4·32
13	1·77	2·16	2·65	3·01	4·22
14	1·76	2·15	2·62	2·98	4·14
15	1·75	2·13	2·60	2·95	4·07
16	1·75	2·12	2·58	2·92	4·02
17	1·74	2·11	2·57	2·90	3·97
18	1·73	2·10	2·55	2·88	3·92
19	1·73	2·09	2·54	2·86	3·88
20	1·73	2·09	2·53	2·85	3·85
21	1·72	2·08	2·52	2·83	3·82
22	1·72	2·07	2·51	2·82	3·79
23	1·71	2·07	2·50	2·81	3·77
24	1·71	2·06	2·49	2·80	3·75
25	1·71	2·06	2·48	2·79	3·73
26	1·71	2·06	2·48	2·78	3·71
27	1·70	2·05	2·47	2·77	3·69
28	1·70	2·05	2·47	2·76	3·67
29	1·70	2·04	2·46	2·76	3·66
30	1·70	2·04	2·46	2·75	3·65
40	1·68	2·02	2·42	2·70	3·55
60	1·67	2·00	2·39	2·66	3·46
120	1·66	1·98	2·36	2·62	3·37
∞	1·65	1·96	2·33	2·58	3·29

Table F is abridged from Table III of Fisher and Yates: 'Statistical Tables for Biological, Agricultural and Medical Research' published by Oliver and Boyd Ltd., Edinburgh, and reproduced by permission of the Authors and Publishers.

TABLE G1. VARIANCE RATIO

	0·20 Significance level								
N_2 \ N_1	1	2	3	4	5	6	12	24	∞
1	9·5	12·0	13·1	13·7	14·0	14·3	14·9	15·2	15·6
2	3·6	4·0	4·2	4·2	4·3	4·3	4·4	4·4	4·5
3	2·7	2·9	2·9	3·0	3·0	3·0	3·0	3·0	3·0
4	2·4	2·5	2·5	2·5	2·5	2·5	2·5	2·4	2·4
5	2·2	2·3	2·3	2·2	2·2	2·2	2·2	2·2	2·1
6	2·1	2·1	2·1	2·1	2·1	2·1	2·0	2·0	2·0
7	2·0	2·0	2·0	2·0	2·0	2·0	1·9	1·9	1·8
8	2·0	2·0	2·0	1·9	1·9	1·9	1·8	1·8	1·7
9	1·9	1·9	1·9	1·9	1·9	1·8	1·7	1·7	1·7
10	1·9	1·9	1·9	1·8	1·8	1·8	1·7	1·7	1·6
11	1·9	1·9	1·8	1·8	1·8	1·8	1·7	1·6	1·6
12	1·8	1·8	1·8	1·8	1·7	1·7	1·7	1·6	1·5
13	1·8	1·8	1·8	1·8	1·7	1·7	1·6	1·6	1·5
14	1·8	1·8	1·8	1·7	1·7	1·7	1·6	1·6	1·5
15	1·8	1·8	1·8	1·7	1·7	1·7	1·6	1·5	1·5
16	1·8	1·8	1·7	1·7	1·7	1·6	1·6	1·5	1·4
17	1·8	1·8	1·7	1·7	1·7	1·6	1·6	1·5	1·4
18	1·8	1·8	1·7	1·7	1·6	1·6	1·5	1·5	1·4
19	1·8	1·8	1·7	1·7	1·6	1·6	1·5	1·5	1·4
20	1·8	1·8	1·7	1·7	1·6	1·6	1·5	1·5	1·4
22	1·8	1·7	1·7	1·6	1·6	1·6	1·5	1·4	1·4
24	1·7	1·7	1·7	1·6	1·6	1·6	1·5	1·4	1·3
26	1·7	1·7	1·7	1·6	1·6	1·6	1·5	1·4	1·3
28	1·7	1·7	1·7	1·6	1·6	1·6	1·5	1·4	1·3
30	1·7	1·7	1·6	1·6	1·6	1·5	1·5	1·4	1·3
40	1·7	1·7	1·6	1·6	1·5	1·5	1·4	1·4	1·2
60	1·7	1·7	1·6	1·6	1·5	1·5	1·4	1·3	1·2
120	1·7	1·6	1·6	1·5	1·5	1·5	1·4	1·3	1·1
∞	1·6	1·6	1·6	1·5	1·5	1·4	1·3	1·2	1·0

Table G1 is abridged from Table V of Fisher and Yates: 'Statistical Tables for Biological, Agricultural and Medical Research' published by Oliver and Boyd Ltd., Edinburgh, and reproduced by permission of the authors and publishers.

TABLE G2. VARIANCE RATIO

	0·05 Significance level								
N_2 \ N_1	1	2	3	4	5	6	12	24	∞
1	161·4	199·5	215·7	224·6	230·2	234·0	243·9	249·0	254·3
2	18·5	19·0	19·2	19·3	19·3	19·3	19·4	19·5	19·5
3	10·1	9·6	9·3	9·1	9·0	8·9	8·7	8·6	8·5
4	7·7	6·9	6·6	6·4	6·3	6·2	5·9	5·8	5·6
5	6·6	5·8	5·4	5·2	5·1	5·0	4·7	4·5	4·4
6	6·0	5·1	4·8	4·5	4·4	4·3	4·0	3·8	3·7
7	5·6	4·7	4·4	4·1	4·0	3·9	3·6	3·4	3·2
8	5·3	4·5	4·1	3·8	3·7	3·6	3·3	3·1	2·9
9	5·1	4·3	3·9	3·6	3·5	3·4	3·1	2·9	2·7
10	5·0	4·1	3·7	3·5	3·3	3·2	2·9	2·7	2·5
11	4·8	4·0	3·6	3·4	3·2	3·1	2·8	2·6	2·4
12	4·8	3·9	3·5	3·3	3·1	3·0	2·7	2·5	2·3
13	4·7	3·8	3·4	3·2	3·0	2·9	2·6	2·4	2·2
14	4·6	3·7	3·3	3·1	3·0	2·9	2·5	2·3	2·1
15	4·5	3·7	3·3	3·1	2·9	2·8	2·5	2·3	2·1
16	4·5	3·6	3·2	3·0	2·9	2·7	2·4	2·2	2·0
17	4·5	3.6	3·2	3·0	2·8	2·7	2·4	2·2	2·0
18	4·4	3·6	3·2	2·9	2·8	2·7	2·3	2·1	1·9
19	4·4	3·5	3·1	2·9	2·7	2·6	2·3	2·1	1·9
20	4·4	3·5	3·1	2·9	2·7	2·6	2·3	2·1	1·8
22	4·3	3·4	3·1	2·8	2·7	2·6	2·2	2·0	1·8
24	4·3	3·4	3·0	2·8	2·6	2·5	2·2	2·0	1·7
26	4·2	3·4	3·0	2·7	2·6	2·5	2·2	2·0	1·7
28	4·2	3·3	3·0	2·7	2·6	2·4	2·1	1·9	1·7
30	4·2	3·3	2·9	2·7	2·5	2·4	2·1	1·9	1·6
40	4·1	3·2	2·8	2·6	2·5	2·3	2·0	1·8	1·5
60	4·0	3·2	2·8	2·5	2·4	2·3	1·9	1·7	1·4
120	3·9	3·1	2·7	2·5	2·3	2·2	1·8	1·6	1·3
∞	3·8	3·0	2·6	2·4	2·2	2·1	1·8	1·5	1·0

Table G2 is abridged from Table V of Fisher and Yates: 'Statistical Tables for Biological, Agricultural and Medical Research' published by Oliver and Boyd Ltd., Edinburgh, and reproduced by permission of the authors and publishers.

TABLE G3. VARIANCE RATIO

	0·01 Significance level									
N_2 \ N_1	1	2	3	4	5	6	8	12	24	∞
1	4052	4999	5403	5625	5764	5859	5981	6106	6234	6366
2	98·5	99·0	99·2	99·3	99·3	99·3	99·3	99·4	99·5	99·5
3	34·1	30·8	29·5	28·7	28·2	27·9	27·5	27·1	26·6	26·1
4	21·2	18·0	16·7	16·0	15·5	15·2	14·8	14·4	13·9	13·5
5	16·3	13·3	12·1	11·4	11·0	10·7	10·3	9·9	9·5	9·0
6	13·7	10·9	9·8	9·2	8·8	8·5	8·1	7·7	7·3	6·9
7	12·3	9·6	8·5	7·9	7·5	7·2	6·8	6·5	6·1	5·7
8	11·3	8·7	7·6	7·0	6·6	6·4	6·0	5·7	5·3	4·9
9	10·6	8·0	7·0	6·4	6·1	5·8	5·5	5·1	4·7	4·3
10	10·0	7·6	6·6	6·0	5·6	5·4	5·1	4·7	4·3	3·9
11	9·7	7·2	6·2	5·7	5·3	5·1	4·7	4·4	4·0	3·6
12	9·3	6·9	6·0	5·4	5·1	4·8	4·5	4·2	3·8	3·4
13	9·1	6·7	5·7	5·2	4·9	4·6	4·3	4·0	3·6	3·2
14	8·9	6·5	5·6	5·0	4·7	4·5	4·1	3·8	3·4	3·0
15	8·7	6·4	5·4	4·9	4·6	4·3	4·0	3·7	3·3	2·9
16	8·5	6·2	5·3	4·8	4·4	4·2	3·9	3·6	3·2	2·8
17	8·4	6·1	5·2	4·7	4·3	4·1	3·8	3·5	3·1	2·7
18	8·3	6·0	5·1	4·6	4·3	4·0	3·7	3·4	3·0	2·6
19	8·2	5·9	5·0	4·5	4·2	3·9	3·6	3·3	2·9	2·5
20	8·1	5·9	4·9	4·4	4·1	3·9	3·6	3·2	2·9	2·4
22	7·9	5·7	4·8	4·3	4·0	3·8	3·5	3·1	2·8	2·3
24	7·8	5·6	4·7	4·2	3·9	3·7	3·3	3·0	2·7	2·2
26	7·7	5·5	4·6	4·1	3·8	3·6	3·3	3·0	2·6	2·1
28	7·6	5·5	4·6	4·1	3·8	3·5	3·2	2·9	2·5	2·1
30	7·6	5·4	4·5	4·0	3·7	3·5	3·2	2·8	2·5	2·0
40	7·3	5·2	4·3	3·8	3·5	3·3	3·0	2·7	2·3	1·8
60	7·1	5·0	4·1	3·7	3·3	3·1	2·8	2·5	2·1	1·6
120	6·9	4·8	4·0	3·5	3·2	3·0	2·7	2·3	2·0	1·4
∞	6·6	4·6	3·8	3·3	3·0	2·8	2·5	2·2	1·8	1·0

Table G3 is abridged from Table V of Fisher and Yates: 'Statistical Tables for Biological, Agricultural and Medical Research' published by Oliver and Boyd Ltd., Edinburgh, and reproduced by permission of the authors and publishers.

TABLE H. VALUES OF χ^2

Degrees of freedom	Probability									
	0·99	0·98	0·95	0·90	0·50	0·10	0·05	0·02	0·01	0·001
1	0·000	0·001	0·004	0·015	0·455	2·71	3·84	5·41	6·64	10·83
2	0·020	0·040	0·103	0·211	1·386	4·61	5·99	7·82	9·21	13·82
3	0·115	0·185	0·352	0·584	2·366	6·25	7·82	9·84	11·34	16·27
4	0·297	0·429	0·711	1·064	3·357	7·78	9·49	11·67	13·28	18·47
5	0·554	0·752	1·145	1·610	4·351	9·24	11·07	13·39	15·09	20·52
6	0·872	1·134	1·635	2·204	5·35	10·65	12·59	15·03	16·81	22·46
7	1·239	1·564	2·167	2·833	6·35	12·02	14·07	16·62	18·48	24·32
8	1·646	2·032	2·733	3·490	7·34	13·36	15·51	18·17	20·09	26·13
9	2·088	2·532	3·325	4·168	8·34	14·68	16·92	19·68	21·67	27·88
10	2·558	3·059	3·940	4·865	9·34	15·99	18·31	21·16	23·21	29·59
11	3·05	3·61	4·57	5·58	10·34	17·28	19·68	22·62	24·73	31·26
12	3·57	4·18	5·23	6·30	11·34	18·55	21·03	24·05	26·22	32·91
13	4·11	4·76	5·89	7·04	12·34	19·81	22·36	25·47	27·69	34·53
14	4·66	5·37	6·57	7·79	13·34	21·06	23·69	26·87	29·14	36·12
15	5·23	5·99	7·26	8·55	14·34	22·31	25·00	28·26	30·58	37·70
16	5·81	6·61	7·96	9·31	15·34	23·54	26·30	29·63	32·00	39·25
17	6·41	7·26	8·67	10·09	16·34	24·77	27·59	31·00	33·41	40·79
18	7·02	7·91	9·39	10·87	17·34	25·99	28·87	32 35	34·81	42·31
19	7·63	8·57	10·12	11·65	18·34	27·20	30·14	33·69	36·19	43·82
20	8·26	9·24	10·85	12·44	19·34	28·41	31·41	35·02	37·57	45·32
21	8·90	9·91	11·59	13·24	20·34	29·61	32·67	36·34	38·93	46·80
22	9·54	10·60	12·34	14·04	21·34	30·81	33·92	37·66	40·29	48·27
23	10·20	11·29	13·09	14·85	22·34	32·01	35·17	38·97	41·64	49·73
24	10·86	11·99	13·85	15·66	23·34	33·20	36·42	40·27	42·98	51·18
25	11·52	12·70	14·61	16·47	24·34	34·38	37·65	41·57	44·31	52·62
26	12·20	13·41	15·38	17·29	25·34	35·56	38·89	42·86	45·64	54·05
27	12·88	14·12	16·15	18·11	26·34	36·74	40·11	44·14	46·96	55·48
28	13·56	14·85	16·93	18·94	27·34	37·92	41·34	45·42	48·28	56·89
29	14·26	15·57	17·71	19·77	28·34	39·09	42·56	46·69	49·59	58·30
30	14·95	16·31	18·49	20·60	29·34	40·26	43·77	47·96	50·89	59·70

For large values of the degrees of freedom the expression $\sqrt{2\chi^2} - \sqrt{(2n-1)}$ may be used as a normal deviate with unit variance (Table E).

Table H is abridged from Table IV of Fisher and Yates: 'Statistical Tables for Biological, Agricultural and Medical Research,' published by Oliver and Boyd Ltd., Edinburgh, and reproduced by permission of the authors and publishers.

BIBLIOGRAPHY

The numbered references are those referred to in the text.

1. Moroney, M.J.: *Facts from Figures.* Penguin Books (London). An extensive bibliography is given in this book.
2. Quenouille, M.H.: *Rapid Statistical Calculations.* Griffin (London). A small book giving a number of quick, though sometimes approximate, methods of doing statistical calculations.
3. Huff, D.: *How to Lie with Statistics.* Gollancz (London). A light and humorously written book illustrating the base uses to which statistics can be put, either by ignorance or malice, or both! A book all should read.
4. Mather, K.: *Statistical Analysis in Biology.* Methuen (London). Describes the theory of many of the processes of statistical analysis very fully, but all the examples are taken from biology.
5. Tippett, L. H. C.: *Statistics.* Oxford University Press (London). A 'popular' type of book which deals more especially with sampling, and the general principles.
6. Fisher, R. A.: *Statistical Methods for Research Workers.* Oliver & Boyd (Edinburgh). One of the classical works on the subject, which has run into many editions. It is not, however, intended for beginners.
7. Lindley, D. V., and Miller, J. C. P.: *Cambridge Elementary Statistical Tables.* Cambridge University Press. A useful and inexpensive set of tables.
   Fisher, R. A., and Yates, F.: *Statistical Tables for Biological Agricultural and Medical Research.* Oliver & Boyd (Edinburgh).

Other Works

Brownlee, K. A.: *Industrial Experimentation.* H.M.S.O. (London).

Davies, O. L.: *Statistical Methods in Research and Production.* Oliver & Boyd (Edinburgh).

Deming, W. E.: *Some Theory of Sampling.* Wiley (New York).

Gerlough, D. L., and Schuhl, A.: *Poisson and Traffic.* The Eno Foundation (Saugatuck, U.S.A.).

Goulden, C. H.: *Methods of Statistical Analysis.* Wiley (New York).

Greenshields, B. D., and Weida, F. M.: *Statistics with Application to Highway Traffic Analyses.* The Eno Foundation (Saugatuck, U.S.A.).

Halstead, H. J.: *An Introduction to Statistical Methods.* Macmillan (Melbourne).

Votaw, D. F. (Jr.) and Levison, H. F.: *Elementary Sampling for Traffic Engineers.* The Eno Foundation (Saugatuck, U.S.A.).

Yule, G. U., and Kendall, M. G.: *An Introduction to the Theory of Statistics.* Griffin (London). In spite of the title this is a very comprehensive work.

ANSWERS TO EXAMPLES

Example No.	*Answer*
2.1	Diagrams are shown in Fig. 15. (Note spread of distribution in Fig. 15a.) 5% of years above 22 in per winter 16% of years below 12 in per winter

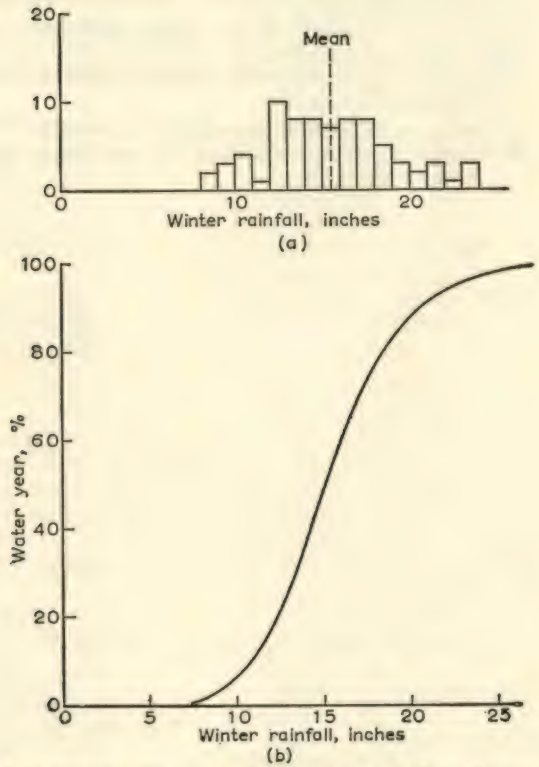


Fig. 15. Histogram and percentile curves plotted from Example 2.1

4.1	Mean for Connecticut: 3·49 deaths per 10^8 vehicle-miles Mean for Rhode Island: 3·02 deaths per 10^8 vehicle-miles
4.2	Mean annual winter rainfall 1883–1959: 15·50 in

5.1 *Connecticut* $s^2 = 0{\cdot}412$ $s = 0{\cdot}642$
Rhode Island $s^2 = 0{\cdot}415$ $s = 0{\cdot}644$
(Both without Sheppard's adjustment)

5.2 $s^2 = 13{\cdot}12$ $s = 3{\cdot}62$, without Sheppard's adjustment

5.3 $\bar{F} = 4740$ cusec $s = 1312$ F_m is between 5100 and 6300 cusec

8.1 $p = 0{\cdot}02$. Rhode Island rate is significantly lower than that of Connecticut.

8.2 Mean squares 1948 = 7·36 and 1949 = 4·14, are significantly different. Use modified method. $0{\cdot}9 > p > 0{\cdot}8$. Means definitely not significantly different

9.1 $p > 0{\cdot}20$. Times not significantly different

9.2 Mean square 'between arrays' is highly significantly larger than that within arrays. 'Between arrays' of same laboratory is not significantly different. Observers C and X are significantly different from rest of observers. They were chosen because they had little experience with the experimental apparatus. The analysis has shown this

10.1 *Connecticut* $Y = 4{\cdot}47 - 0{\cdot}14x$ Highly significant, $p < 0{\cdot}001$
Rhode Island $Y = 3{\cdot}75 - 0{\cdot}10x$ Significant, $0{\cdot}05 > p > 0{\cdot}02$
Both zeros in 1946. Difference between regression coefficients not significant, $0{\cdot}5 > p > 0{\cdot}3$

12.1 (*a*) Reduction in accidents not significant, $p = 0{\cdot}20$
(*b*) Severity of accidents decidedly not significantly reduced, $0{\cdot}9 > p > 0{\cdot}8$

INDEX

The numbers in italics refer to pages on which terms are first introduced and defined or explained

LEEMING STATISTICAL METHODS FOR ENGINEE 519.5 024 62 LEE BLACKIE